



UNIVERSITE
JEAN MOULIN
LYON 3

15, Quai Claude Bernard - 69 007 LYON

FACULTE DES LETTRES ET CIVILISATIONS

Département des sciences de l'information
et de la documentation

Traitement de publications CERN de l'intranet

Importation automatique / semi-automatique de publications
d'expériences CERN dans le catalogue de la bibliothèque.

Travail d'Etude et de Recherche pour la Maîtrise en Documentation
Rapport de stage au CERN du 15 mai au 25 septembre 1999



Présenté par :
Philippe RICANET

Année universitaire:
1999

Directeur de recherche :
M^{me} Jocelyne MILAN
Professeur à l'université
Jean Moulin Lyon III

Responsable de stage :
M^{me} Ingrid
GERETSCHLAGER
Conservateur Enssib,
Responsable du service
des prêtirages CERN.

Remerciements

Je voudrais remercier toutes les personnes qui m'ont soutenu au cours de ce stage de fin d'étude au sein du service d'information scientifique du CERN.

- Madame Ingrid Geretschläger et Monsieur Corrado Pettenati pour m'avoir permis d'effectuer ce stage et accordé leur confiance.

- Catherine Cart et Jocelyne Jerdelet pour leur soutien, leur aide précieuse, leur patience et leur gentillesse.

- tous les membres du service sans exception pour m'avoir si bien accepté dans leur équipe et leur disponibilité.

Merci à Gérard et Martine, mes hôtes pendant 5 mois, sans qui ce stage n'aurait pas été possible.

Traitement de publications CERN de l'intranet

Importation automatique / semi-automatique de publications d'expériences CERN
dans le catalogue de la bibliothèque.

Table des matières.

1. PRÉSENTATION DU CONTEXTE DU STAGE.....	5
1.1 PRÉSENTATION DU CERN.....	5
1.1.1 Aspect historique et développement.....	6
1.1.2 La structure hiérarchique au CERN.....	10
1.2 LA BIBLIOTHÈQUE : LE SERVICE D'INFORMATION SCIENTIFIQUE.....	11
1.2.1 La section de gestion des documents.....	11
1.2.2 Origine des programmes de traitement des données bibliographiques.....	14
1.3 LE SGBD DU CERN ET INTERNET.....	14
1.3.1 Présentation du logiciel de GBD et de gestion de bibliothèque du CERN : Aleph 300	14
1.3.2 Liens hypertextes avec Aleph.....	16
1.3.3 Le WEB.....	16
2. IMPORTATION AUTOMATIQUE / SEMI-AUTOMATIQUE DES PUBLICATIONS DE L'INTRANET DANS ALEPH.....	17
2.1 ANALYSE.....	18
2.1.1 Analyse de la demande et reformulation.....	18
2.1.2 Analyse des publications dans l'intranet.....	19
2.1.3 Recherche des sources de publications sur les expériences CERN.....	27
2.1.4 Analyse du nombre de sites d'expériences.....	29
2.2 PHASE DE DÉVELOPPEMENT.....	34
2.2.1 Exploitation des sources par un listing systématisé en fonction des besoins.....	34
2.2.2 Méthodologie d'une recherche exhaustive sur internet.....	35
2.2.3 Analyse des résultats obtenus.....	38
2.3 RECHERCHE D'UNE MÉTHODE AUTOMATISÉE D'IMPORTATION DES DONNÉES PUIS, MISE EN PLACE D'UN SYSTÈME BASIQUE DE VEILLE DOCUMENTAIRE.....	39
2.3.1 L'importation automatique de publication CERN de l'intranet.....	39
2.3.2 Une aide à la publication sur internet.....	40
2.4 PHASE D'ABOUTISSEMENT.....	49
<i>Sur ces 5 solutions trouvées, seules quatre furent sélectionnées.</i>	49
2.4.1 Résultats obtenus lors de la phase de développement.....	49
2.4.2 Comment envoyer les pages de destinations dans les différents répertoires du site WEB distant	52
3. LES AUTRES ACTIVITÉS LORS DE CE STAGE AU CERN.....	60
3.1 L'UTILISATION PRATIQUE D'ALEPH :.....	60
3.1.1 La recherche d'une notice bibliographique peut être effectuée selon trois modes différents.....	61
3.1.2 Le catalogage :.....	63
3.1.3 Création de renvois sous aleph 300.....	63
3.2 APPRENTISSAGE D'UNIX, DU LATEX, ET D'EMACS.....	64
3.2.1 Unix.....	64
3.2.2 Emacs.....	66
3.2.3 Le LATEX (prononcez LATEK).....	67
3.3 AGIV : UN EXEMPLE DE TRAITEMENT SEMI-AUTOMATISÉ DES AUTEURS.....	68
3.4 ELABORATION D'UN PETIT DOSSIER SUR LES MÉTAMOTEURS POUR LA PRÉSENTATION DE COPERNIC99.....	68

0 Introduction

Ce mémoire présente un travail de stage en maîtrise des sciences de l'information et de la documentation pour l'université Jean Moulin Lyon III. Il a été réalisé au Laboratoire Européen pour la Physique des Particules sous la direction de madame Ingrid Geretschläger, conservateur ENSSIB et responsable de la section du traitement documentaire des prétrirages.

Le CERN est l'un des centres de recherche en physique des particules les plus renommés. Il met en œuvre plus de 150 expériences propres. Au total, avec les celles de moindre importance réalisées en collaboration avec plus de 650 universités du monde entier, on peut évaluer ce chiffre à plus de 800.

Ma mission au sein de la bibliothèque du CERN a été de proposer une solution, voire un processus de veille, afin d'automatiser le rapatriement des publications de chercheurs. Ceux-ci, pris par leurs travaux de haut niveau scientifique, omettent de soumettre leurs publications au CERN mais diffusent leurs recherches dans l'intranet du CERN (ou directement sur internet, comme nous le verrons dans le développement).

Articulé en trois parties, ce mémoire veut dégager dans un premier temps le contexte de ce stage. Ensuite, il aborde de façon plus approfondie la partie réflexive et méthodologique de ce travail au CERN. Enfin, sont traitées les diverses activités annexes moins théoriques et plus pratiques que j'ai eu à effectuer au cours de cette période de stage au CERN.

1. Présentation du contexte du stage.

1.1 Présentation du CERN

A première vue, le CERN donne l'impression d'une petite ville industrielle aux bâtiments désordonnés, articulés autour des machines de production des particules à haute énergie. Le CERN dont l'acronyme actuel est Laboratoire Européen pour la Physique des Particules était appelé autrefois Conseil Européen

pour la Recherche Nucléaire. Depuis 1954, il étudie les particules fondamentales qui composent la matière et les forces qui les unissent.

Ses objectifs définis sont : faire progresser les connaissances en physique des particules à des fins non militaires, développer et attirer en Europe les collaborations scientifiques en physique nucléaire et encourager les progrès techniques en rendant publics régulièrement tous ses travaux.

1.1.1 Aspect historique et développement

1.1.1.1 La création du CERN

Après 1945, l'Europe scientifique sort exsangue de la seconde guerre mondiale. L'hémorragie des grands cerveaux européens commencée avant la guerre se poursuit tant à cause de problèmes de financement des laboratoires que pour des raisons politiques. L'esprit européen naissant du début des années 50 prend conscience de sa force financière et scientifique et choisit l'union des nations dans le but de créer des laboratoires capables de rivaliser en tous points avec les américains.

Une organisation provisoire CERN émerge en 1952 du regroupement de l'Allemagne, la Belgique, le Danemark, la France, la Grèce, l'Italie, la Norvège, les Pays-Bas, la Suède, la Suisse et la Yougoslavie. Sa convention est signée en 1953 puis ratifiée en 1954 par 9 pays.

En 1955, le CERN s'agrandit de 4 états membres et devient l'organisation européenne pour la recherche nucléaire. En 1959, l'Autriche adhère, suivie de l'Espagne en 1961 qui remplace la Yougoslavie qui se retire à cette date. Quant à la Finlande, elle adhère en 1991.

Les pays collaborant au CERN possèdent 3 statuts différents :

20 pays membres nommés ci-dessus financent le CERN proportionnellement à leur PIB et 35 états non-membres utilisent les structures et s'autofinancent. L'ensemble représente 1705 chercheurs et 211 universités. Les pays observateurs tels que les USA, la Russie, Israël, le Japon, la Turquie et les membres de l'Unesco.

1.1.1.2 Le Programme scientifique

1.1.1.2.1 De 1957 à 1983 :

En 1957, naît le synchrocyclotron de 600 mégaélectronvolt (MeV) permettant aux scientifiques d'observer la désintégration d'un pion en un électron et un neutrino.

Le P.S. (synchrotron à protons) est mis en service en 1959. Cet accélérateur, avec 28 GeV (gigaélectronvolt) est alors la plus puissante machine du monde. C'est en 1963 que l'on inaugure ce qui deviendra une spécialité du programme d'expérimentation du **CERN** : les expériences avec des faisceaux de neutrinos. La première machine au monde à produire des collisions frontales entre protons, appelée les **ISR** débute ses expérimentations en 1971 et s'arrête en 1984. Ce furent deux anneaux magnétiques d'une circonférence d'environ 1 kilomètre chacun dans lesquels les protons accélérés à la vitesse de la lumière se collisionnent de front et interagissent en révélant ainsi la composition de la matière. Quarks, Neutrinos, et divers composants de matière et d'antimatière jaillissent de ces collisions.

1973 : construction de la grande chambre à bulles européenne (**B.E.B.C.** : *Big European Bubble Chamber*) cette machine posséda le plus grand aimant supraconducteur du monde. Elle prit jusqu'en 1984 environ 6 300 000 clichés photographiques de grande valeur tant scientifique qu'esthétique pour sa ressemblance avec des œuvres d'art abstrait.

1976 : création du S.P.S (*superproton synchrotron*) : Cet accélérateur à cibles fixes de taille pharaonique, enterré à 50 mètres sous terre et au diamètre de 7 kilomètres est transformé à partir de 1981 en collisionneur proton-antiproton. La première collision proton-antiproton est observée en juillet 1981. En découle en 1983, la découverte de nouvelles particules, les bosons W à et le boson Z 0. Le succès scientifique de ce type de collision grâce aux deux éminents cernois Carlo Rubbia et Simon van der Meer est couronné par un double prix Nobel de physique en 1984.

1.1.1.2.2 Après 1983 : le LEP.

Fort de ses succès, le CERN fait encore plus gros, encore plus fort avec le grand collisionneur à électrons-positons dit **LEP** (*large electron-positron collider*). Le LEP mobilise, par ses expériences en collaboration internationale, des milliers d'ingénieurs et de chercheurs. Construite de 1983 à 1989 c'est l'une des machines les plus complexe et les plus grandes au monde possédant 50 000 pièces, 5388 aimants, les 27 kilomètres de circonférence du SPS, un diamètre intérieur de 3,8 mètres, et un aimant supraconducteur de 7 500 tonnes. L'ensemble forme un anneau enterré de 50 à 170 mètres sous terre permettant d'obtenir des collisions frontales de particules accélérées à la vitesse de la lumière. Dès 1982, des spécialistes en physique des particules proposèrent des sujets d'expériences à réaliser avec le LEP.

Les 4 expériences LEP approuvées par le CERN sont **Aleph**, **Delphi**, **L3** et **Opal**.

Avant la création du LEP, en 1989, seules quelques centaines de bosons Z0 avaient été observées de par le monde. Le but du LEP était d'en créer un maximum afin de pouvoir les observer. Dès la première année, il en produisit : son objectif était atteint. Les plus importants résultats de ces expériences sont : le calcul de la durée de vie du boson Z0 et la déduction de l'existence de seulement trois familles de particules de matière dans l'univers.

Dans le milieu des années 80, une série de discussions à eu lieu au CERN pour poursuivre les expériences LEP avec le LHC (*Large Hadron Collider*).

1.1.1.2.3 Depuis 1994 : le LHC

En 1994, le Conseil du CERN a approuvé la construction du **LHC**, dont la mise en service est prévue pour 2005. Le «grand collisionneur de hadrons», mettra des protons en collision frontale à des énergies dix fois supérieures aux énergies disponibles à ce jour dans le monde (8 000 milliards d'électronvolts) afin de pénétrer encore plus profondément dans la matière et d'approcher les conditions originales de l'univers (c'est à dire les 10^{-12} secondes qui ont suivi le "*Big-Bang*"). Comme le LEP, quatre expériences LHC ont été approuvées. **ATLAS** (*A Toroidal LHC Apparatus*) qui est une collaboration de 150 instituts dont le but premier est la compréhension de

la nature des masses, **CMS** (*Compact Muon Solenoid*), **ALICE** (*A Large Ion Collider Experiment*) qui étudiera les collisions frontales de neutrons à des énergies très élevées afin d'observer une nouvelle phase de la matière nommée le plasma quark-gluon, enfin, **LHCb** qui servira à l'étude de la désintégration de particules nommées Bosons B au LHC grâce à l'instrument de détection de matière et d'antimatière le plus sensible au monde.

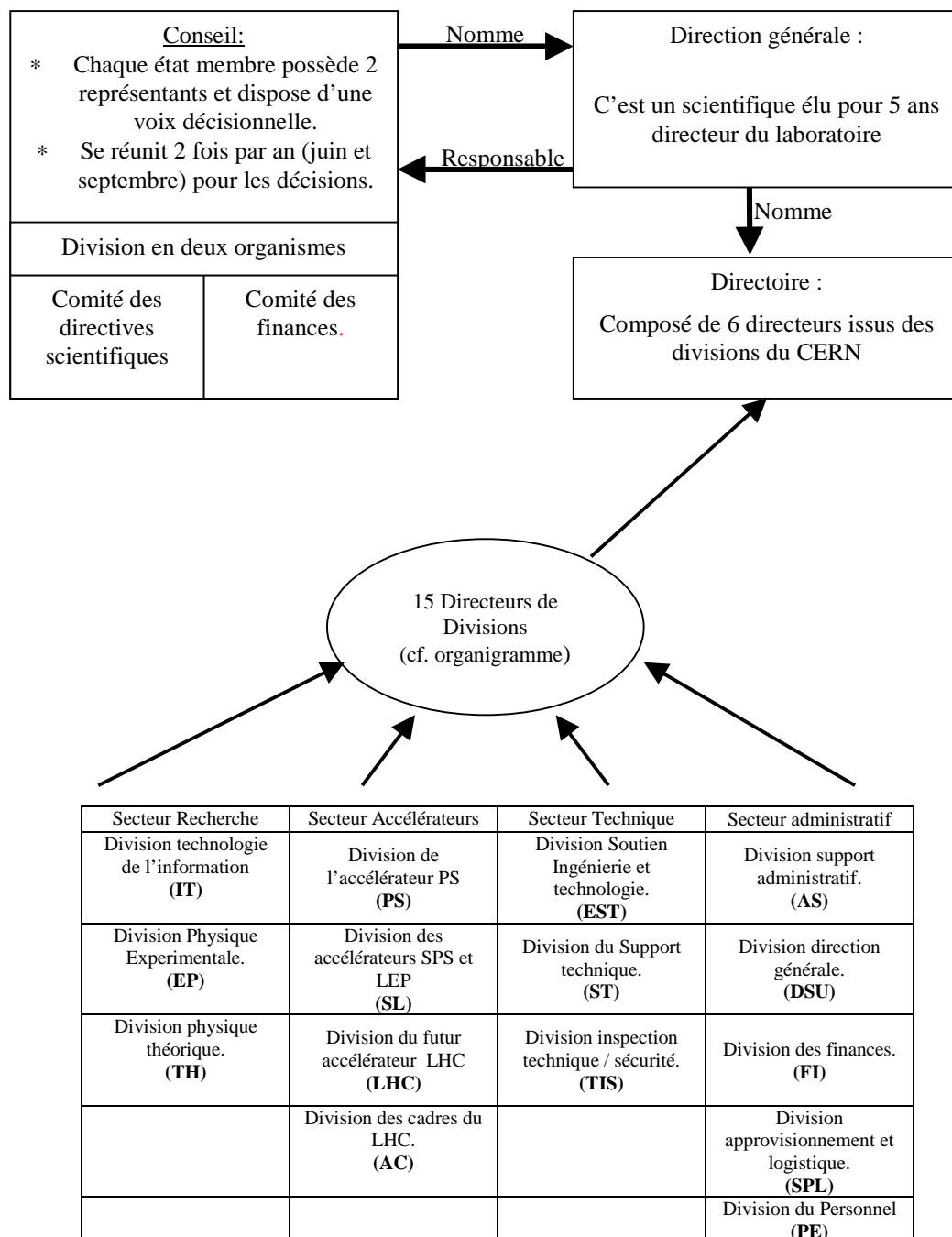
1.1.1.3 Le CERN : une réussite certaine

Le CERN est actuellement le plus grand laboratoire européen de recherche fondamentale. Son site couvre 559 hectares entre Suisse et France, regroupe plus de 6 500 chercheurs et ingénieurs, et 2500 membres du personnel. En 1998, son budget annuel s'élève à 950 millions de francs suisses, répartis entre les états membres selon leur revenu national. Le CERN a inspiré la création d'autres instituts européens dont : la Société Européenne de Physique, l'Organisation Européenne de Biologie Moléculaire, l'Organisation Européen de Biologie Moléculaire et l'Organisation européenne pour la recherche astronomique dans l'hémisphère austral (E.S.O.). Les accélérateurs, dont ceux du CERN, sont devenus indispensables à l'humanité pour comprendre ce qui s'est passé il y a 15 milliards d'années, au moment du *Big-Bang*.

Le CERN est grand consommateur d'appareillage de haute technologie spécialisée en électronique, optique, cryogénie, supra-conductivité, hautes fréquences, etc. ...Pour faire face à cet appétit, il dut développer toutes ces technologies qui, par transfert dans le domaine public, profitent à l'humanité entière. Nommons a titre d'exemple les progrès médicaux liés à la thérapie à hadrons (les rayons dans le traitement contre le cancer) ou à l'utilisation du scanner médical...

La CERN risque de fonctionner encore longtemps car la physique des particules reste une science nouvelle et un grand champ de recherche à défricher car seules les particules les plus élémentaires ont été découvertes.

1.1.2 La structure hiérarchique au CERN



Le CERN est une organisation internationale fortement hiérarchisée. Elle possède deux pouvoirs décisionnels à la tête du laboratoire. Le conseil et un directeur élu pour 4 ans. Le directeur, actuellement monsieur Mayani, est assisté dans ses

prises de décisions par un directoire composé 6 directeurs de division sur les 15 existants.

1.2 La bibliothèque : le service d'information scientifique

1.2.1 La section de gestion des documents

1.2.1.1 Présentation de la bibliothèque

La bibliothèque du CERN se compose d'une bibliothèque centrale, de 4 bibliothèques satellites et des archives historiques et scientifiques. Ces bibliothèques sont ouvertes 24 heures sur 24, toute l'année. Seul le *DESK*, le service d'information aux lecteurs de la bibliothèque centrale fait exception.

1.2.1.2 Composition du service.

Le service d'information scientifique est divisé en 4 sections. Elle compte 5 personnes dont la responsable, Ingrid Geretschläger. Elle accueille régulièrement des stagiaires, étudiants et apprentis de différentes écoles et universités internationales. (cf. annexe 1)

1.2.1.2.1 La section gestion des documents (cf. annexe1)

C'est le lieu de ce stage Les différentes activités effectuées par la bibliothèque sont l'acquisition, le traitement et la diffusion documentaire sous toutes ses formes. Elle gère différents documents, tant CERN que non CERN. C'est à dire : les monographies, les prêtirages, les articles, les thèses, les rapports, les métadonnées... Les travaux essentiels de la bibliothèque sont le catalogage, la gestion de base de données, le dépouillement, et l'importation de notices bibliographiques. Elle coopère avec d'autres bibliothèques de recherche telles que SLAC (*Standard Linear accelerator of California*), entre autres.

La section gestion des documents est aussi appelé couramment service des prêtirages, dit, des *pre-prints*. Ceci par ce qu'elle gère quotidiennement un nombre

énorme de prêtirages, soit 600 par semaines. Ce est alimenté principalement par trois flux provenant principalement du CERN, de l'archivage électronique de Los Alamos et du courrier papier. Un prêtirage correspond aux documents de littérature grise écrits par les physiciens et soumis à un journal dans le but d'être publié. Les chercheurs sont particulièrement attirés par ce genre de littérature car un article met 6 mois à 2 ans avant d'être publié. Les chercheurs se fournissent directement prêtirages dans les bases de données de la bibliothèque via internet afin d'obtenir des informations de dernière minute pour leurs études. Ici, il n'y a pas de DSI¹ mais l'exacte procédure inverse. A cet effet, le service des prêtirages intègre au plus vite dans les base de données d'Aleph ces documents non encore publiés .En moyenne, 45% des pré-publications sont importés par les divisions de recherche, 40% par les divisions des accélérateurs et les 15% par le reste du CERN

Ces articles, avec l'ensemble des autres documents CERN font l'objet d'une publication annuelle : le volume III du rapport annuel du CERN. Ce recueil est une liste de documents CERN et non-CERN que le service envoi aux laboratoires du monde entier, surtout à ceux des pays en voie de développement sous équipés en informatique.

1.2.1.2.2 Les trois autres sections

- La section gestion des périodiques qui s'occupe des revues papier et électroniques. Cette section s'oriente résolument vers une option «tout électronique». Elle compte actuellement environ 500 abonnements a des journaux électroniques sur 1000 titres de périodiques.
- Le service aux utilisateurs qui prend en charge les recherches documentaires complexes, gère les prêts inter-bibliothèques, construit des pages WEB du site de la bibliothèque, distribue les publications CERN...
- La section des archives qui indexe, remet en état et effectue des recherches sur l'ancien fonds documentaire qu'il administre.

¹ Diffusion sélective de l'information

1.2.1.3 Budget, fonds et sources de la bibliothèque

Le budget de la bibliothèque s'élève à près d'un million de francs suisses. La répartition de ce financement est consacrée à 50% aux périodiques, à 20 % aux achats et à la restauration de monographies. Cela ne manque pas d'étonner pour un grand laboratoire de recherche tel que le CERN, car cela représente très peu. (0,1% de son budget).

Au total, la bibliothèque gère 40.000 livres et comptes-rendus de conférences (*proceedings*), 270.000 prétirages et articles stockés dans la base de données générale du CERN et plus de 1000 périodiques dont 500 sous forme électronique.

Tous les documents électroniques sont stockés dans les disques durs du serveur central du CERN. à première vue, un tel système de stockage de l'information sur simple support magnétique paraît risqué, car l'exposition au champ magnétique d'un simple aimant peut détruire toutes les bases de données du CERN. Mais une procédure de sauvegarde (*Back-up*) des données appelée « Misk » est lancée quotidiennement, tous les soirs, sur bande magnétique. Ainsi, en cas d'erreur ou de destruction d'information dans une base de donnée, un simple coup de téléphone au support informatique permet de retrouver dans la bande de la veille l'information perdue.

Les sources de la section de gestion de documents sont la production des chercheurs du CERN, les bibliothèques des grandes expériences internationales de physique des particules élémentaires et les grandes bases de données. (Cf : tableau annexe 2).

Le traitement des documents dans la section AS-SI s'accroît de façon exponentielle car si, en 1994, on recensait 12000 documents traités par an. En 1998, elle en gérait environ 36000. Avec un budget qui diminue d'année en année, et une collection qui s'accroît sans cesse, on peut comprendre que la bibliothèque du CERN cherche des solutions pour maintenir la qualité de son service.

De plus en plus en collaboration avec les universités et les organismes de formation, la bibliothèque cherche à développer ses contacts pour accueillir du personnel étudiant, apprenti ou chômeur tout en participant à leur formation.

Parallèlement à cela, la bibliothèque ne cesse d'automatiser le traitement de ses collections et de numériser ses fonds. Ceci explique que le travail quotidien à la bibliothèque devient de plus en plus qualifié et technique.

1.2.2 Origine des programmes de traitement des données bibliographiques

Le service des prêtirages traite deux familles de documents. Les documents CERN et non-CERN, sous forme électronique et papier. Il y a deux ans, le service des *prêtirages* s'est aperçu, lors d'une étude, que les physiciens, surtout les théoriciens du CERN publient leurs productions les plus récentes sur l'intranet du CERN sans pour autant les envoyer dans les banques de données de la bibliothèque.

Ces documents non soumis sont nommés «*by-passed*». Littéralement, ce terme anglais signifie « document hors-circuit ». Pour récupérer ces documents, le service eut l'idée de développer des outils tels que des programmes d'importations.

Ces derniers peuvent être mis au point par la bibliothèque elle-même, à l'aide d'étudiants en formation universitaire soutenus par le support informatique. Mon stage s'inscrit dans ce cadre d'ouverture à l'importation automatique de documents CERN "*by-passed*" dans la base de données ALEPH.

1.3 Le SGBD₂ du CERN et internet

1.3.1 Présentation du logiciel de GBD et de gestion de bibliothèque du CERN : Aleph 300

1.3.1.1 Présentation générale

Conçu par Ex Libris, une société israélienne spécialisée dans le domaine de l'informatisation de médiathèques, ce système de gestion de bibliothèque permet de gérer des médiathèques de tous types (Bibliothèques Universitaires autrichiennes, Bibliothèques du Vatican, Services de Documentation...). Cependant, selon le

² Système de gestion de base de donnée

distributeur, ce logiciel de bibliothèque est particulièrement orienté vers une clientèle possédant une grosse structure et un réseau interne telle que le CERN.

Aleph est un logiciel qui se structure en différentes bases de données et modules. (Cf. annexes 13).

Il permet la recherche documentaire, le catalogage, le prêt entre différentes bibliothèques, la gestion des fournisseurs, des commandes, des prêts, des adhérents, des périodiques... Un des points forts d'Aleph est qu'il peut gérer les alphabets Latin, Arabe, Hébreu, Cyrillique et Grec ; ceci est très avantageux pour un établissement international comme le CERN. Le prix actuel minimum du logiciel Aleph pour 50 postes est d'environ 700.000 F français.

Le système Aleph a été choisi en France et en Suisse par de nombreux établissements prestigieux comme les écoles Polytechniques de France et de Suisse ainsi que nombreuses établissements universitaires du monde entier.

1.3.1.2 Aleph au CERN

Le CERN, possède la version 300 d'Aleph qui fonctionne sous UNIX principalement (l'interface de la version récente, Aleph 500, est de type Windows). Cela implique l'utilisation du langage LaTeX₃ pour les notices, d'Emacs comme éditeur et traitement de texte... Ceci rend l'utilisation de ce logiciel au CERN peu conviviale et ardue pour un néophyte, même si sa version 300 est très complète. Très complète car modifiée et améliorée par les informaticiens du CERN selon les besoins spécifiques à un grand centre de recherche.

Le passage du CERN à Aleph 500 qui fonctionnera avec Oracle, le Système de Gestion de Base de Donnée, est prévu pour les années futures. Mais aucune date n'est vraiment avancée. Aleph 300 a été tellement configuré pour son utilisation spécifique au CERN que ce changement de version risque d'être assez difficile. Par exemple, il faudrait pour cela que la bibliothèque change sa norme de catalogage car les documentalistes du CERN ont créé leur norme propre en adaptant aux besoins du CERN la Norme ACR2 et le format ISBD₄

³ Formateur de texte utilisé pour la rédaction de formules mathématiques

⁴ (International Standard Bibliographic Description : c'est une des normes internationale présentant les éléments et les symboles graphiques de la description bibliographique d'un document quel qu'il soit)

1.3.2 Liens hypertextes avec Aleph

Les notices bibliographiques possèdent un champ prévu pour activer des liens hypertextes jusqu'au document complet : c'est le champ EXT qui se présente sous la forme d'un URL classique. Il se présente ainsi :

EXT \$\$x [http/www.URL.ch](http://www.URL.ch) \$\$n acces to fulltext document

La plupart des documents plein-texte associés sont des fichiers textes, très souvent compressés (au format Zip pour windows ou Gz voire Tar pour Unix) que l'on trouve sous deux formats : PS (Postscript) et PDF (Portable Document Format). Ces formats d'édition et d'impression ont été tous deux inventés par Adobe : l'un en 1984 et l'autre en 1996. Le PDF tend à remplacer le PS et se transforme peu à peu en standard internet pour le stockage de gros dossiers (de plus de 100 pages) en ligne. Par internet, l'utilisateur peut consulter le catalogue de la bibliothèque et consulter ou télécharger les thèses, rapports de conférences et autres documents CERN pour peu que leur ordinateur possède les programmes adéquats.

En général, les grands centres de documentation boudent Aleph car il ne possède pas de véritable thésaurus, ne se décompose pas en modules même s'il peut fonctionner comme Système de Gestion de Base de Donnée.

1.3.3 Le WEB

Inventé et développé au CERN par Tim Berners-Lee, le WEB a été créé pour faciliter le partage d'informations entre des groupes de chercheurs dispersés sur la planète. Ce sont eux qui créèrent le besoin d'accéder à des bases de données, d'échanger et de créer des documents scientifiques, des articles, des rapports ... à distance et à tout moment.

Internet a été préalablement inventé par le pentagone dans les années 60 comme système de défense des USA et était déjà largement utilisé par les chercheurs. Le CERN développa un logiciel spécifiquement pour eux et son succès fut tel qu'il devint rapidement un outil universel qui démocratisa Internet.

Tim Berners-Lee créa toutes les normes WEB actuelles : le premier serveur WWW, le premier logiciel de communication, la forme des URLs

[Http://www.url.com](http://www.url.com) et le langage HTML. Tout ceci à apporté, par la géniale simplicité, de ce système une norme pour les communications à distance entre ordinateurs. Si l'on souhaite donner une définition succincte du WEB, il faudrait dire que c'est une énorme page HTML, accessible par l'Internet, dans lequel est entreposé le savoir global. Le mode de navigation hypertexte est appelé "butinage", "surf" ou encore "browsing" pour les anglophones. En théorie, le WWW propose librement à tout utilisateur l'ensemble des informations de la planète dont il dispose. Le seul réel problème actuel, lié à son développement semble être la pertinence et la validité de l'information.

Le CERN, actuellement, possède un réseau d'ordinateurs important et complexe de 12.000 systèmes reliés en réseau dont environ 9.000 ordinateurs de bureaux, des bases de fichiers, des centres de calcul et diverses installations spécialisées. Les liens de l'intranet du CERN avec le WEB ont une capacité totale d'environ 10 millions de bits par seconde. Le WEB, né au CERN garde une liaison privilégiée avec son nid.

2. Importation automatique / semi-automatique des publications de l'intranet dans Aleph

L'approche de ce sujet nécessitait de suivre au maximum une méthode scientifique de traitement de l'information telle que celle enseignée au département des sciences de l'information et de la documentation de l'université Jean Moulin Lyon III. J'ai voulu m'y astreindre afin d'espérer obtenir un résultat optimal.

A savoir, théoriquement :

- a- Demande de l'utilisateur
- b- Analyse des besoins
- c- Reformulation de la demande en langage documentaire
- d- Analyse des sources
- e- Interrogation des sources
- f- Extraction et exploitation des résultats
- g- Validation de la réponse (si non, retour à la reformulation de la demande)
- h- Réponse satisfaisante fournie à l'utilisateur

Pratiquement, en raison de modifications de circonstances liées à la conjoncture, la méthodologie s'est retrouvée sensiblement transformée en trois parties que j'ai voulu appliquer à mon plan.

- a- Phase d'analyse
- b- Phase de développement
- c- Phase d'aboutissement

2.1 Analyse

2.1.1 Analyse de la demande et reformulation .

La demande formulée par le section d'Ingrid Geretschläger était la suivante. Il s'agissait de trouver un moyen d'importer automatiquement depuis intranet toutes les pages de publications concernant les expériences du CERN. Je disposais d'une semaine pour me familiariser avec les commandes Unix ainsi que les utilitaires associés, tous les outils et le logiciel Aleph. Ensuite, je devais le plus rapidement possible élaborer une stratégie de recherche et commencer à trouver des solutions pour faciliter l'importation des données diffusées dans l'intranet du CERN par les chercheurs.

L'organisation choisie pour aborder ce travail fut la décomposition des objectifs de ce stage jusqu'à son plus petit élément. L'idée était de décortiquer cette demande en plusieurs objectifs concrets à atteindre afin de mieux appréhender et répondre à la demande. Les différentes étapes redéfinies étaient les suivantes :

- a- Evaluer les besoins à partir d'un site précis
- b- Trouver toutes les différentes expériences et les répertorier
- c- Rechercher la localisation des informations dans chaque site d'expérience et relever leurs URL en vue d'une importation automatique
- d- Trouver une ou des solutions pour importer les informations recueillies
- e- Appliquer les solutions et analyser les résultats obtenus

2.1.2 Analyse des publications dans l'intranet

Afin d'être le plus efficace possible, j'ai dû évaluer la pertinence d'une telle recherche. En effet, il était impensable de se lancer dans une recherche exhaustive des expériences et des collaborations dans intranet sans faire préalablement l'analyse plus ou moins complète des publications d'un site d'expérience type.

Pour recueillir les informations, un tableau à double entrée avec Microsoft Excel 97 m'a semblé être l'outil adéquat car il permettait de tirer assez rapidement des chiffres et des conclusions d'une courte analyse.



Le site choisi pour une telle analyse fut le site de l'expérience L3 pour sa localisation connue dans l'intranet du CERN et pour sa représentativité. L3 a été choisie pour cette étude car c'est une expérience connue comme étant l'une des 4 expériences principales du LEP. Son site est facilement accessible puisqu'il se trouve dans le listing principal du CERN. De plus, ses pages de publication sont aisément localisables. Les publications recueillies dans site L3 sont de 6 grands types ; les

notes internes, les publications, les pré-publications soumises lors d'une conférences, les journaux et les thèses.

En effectuant un sondage dans tous ces types de publications et en évaluant ensuite le taux de présence d'un type de publication dans la base générale, une vision globale du travail à pu ainsi être extraite.

Les conclusions que l'on put tirer de ce tableau sont les suivantes :

Pour le type de documentation 1 : Thèses

	Échantillon-nage	Nombre total de références	Nombre de notices hors d'Aleph	Nombre de notices dans Aleph	% de notices consultées
Hefei	1	1	1	0	100%
Beijing	1	1	0	1	100%
Lyon	1	1	1	0	100%
LAPP	1	1	1	0	100%
Aachen I	1	2	1	0	50%
Aachen III	1	3	1	0	33%
Humboldt Univ. Berlin	2	4	0	2	50%
Desy - Zeuthen	2	2	2	0	100%
Amsterdam	2	3	2	0	67%
Nijmegen	2	2	2	0	100%
Nikhef	3	3	3	0	100%
Utrecht	3	3	3	0	100%
TATA	4	4	4	0	100%
Florence	4	4	4	0	100%
Perugia	5	14	5	0	36%
Rome (Longo)	5	55	4	1	9%
Rome (Dionisi)	5	5	1	4	100%
Kaist	5	5	0	5	100%
ITEP	5	11	0	5	45%
CIEMAT	5	27	5	0	19%
Genève	5	22	5	0	23%
ETH	5	14	5	0	36%
Boston University	5	28	5	0	18%
California	5	31	4	1	16%
Carnegie Mellon	5	6	3	2	83%
Harvard	5	8	5	0	63%
John Hopkins	5	14	5	0	36%
Michigan	6	6	6	0	100%
MIT	7	7	0	7	100%
Northeastern	7	16	4	3	44%
Prinetown	2	3	1	1	67%
TOTAL des Thèses	115	306	83	32	38%

A partir de cet échantillonnage d'environ un tiers effectué sur 306 notices, j'ai noté que seulement 27% des thèses ont été soumises à la bibliothèque.

Cela représente environ 220 documents à saisir au total.

Type de documentation 2 : Publications

	Échantillonnage	Nombre total de références	Nombre de notices hors de la base	Nombre de notices dans la base	% de notices consultées
Publications	174	174	2	172	100%

L'étude portée sur les publications du site principal de L3 a été exhaustive. Sur l'ensemble, à peine 1% ne sont pas recensées. Cela signifie que les documents physiquement publiés sont en général soumis à la bibliothèque du CERN.

Type de documentation : Divers

	Échantillonnage	Nombre total de références	Nombre de notices hors d'Aleph	Nombre de notices dans Aleph	% de notices consultées
Other L3 publication	15	65	9	6	23%

Environ un quart de ces documents ont été analysés. Sur ce petit quart, j'ai noté que 60% (9 sur 15 échantillons) sont absentes de la base Aleph. Au total, cela représenterait 39 notices de publications.

Type de documentation 4 : les contributions aux conférences

	Nom	Echantillonnage	Nombre total de références	Nombre de notices hors de la base	Nombre de notices dans la base	% de notices consultées
L3 papers Contributed to conferences	March: Moriond, Les Arcs, France 1999	21	21	21	0	100%

Abstracts for APS99 Conference 1999	13	13	12	1	100%
March: Moriond, Les Arcs, France 1998	10	45	9	1	22%
March: APS, Atlanta, USA 1999	13	13	12	1	100%
Jerusalem 97 abstract	10	52	6	4	19%
Lepton-Photon Symposium in Hamburg papers	10	27	0	10	37%
Links to papers and write-ups	7	11	4	3	64%
TOTAL	84	182	64	20	46%

Les contributions aux conférences sont à 76% absentes d'Aleph. Au total, cela représente 138 documents absents de la base pour 182 recensés.

Type de documentation 5 : présentations de Conférences

Les présentations de conférences affichent le plus mauvais chiffre obtenu, avec une évaluation de 165 notices manquantes, sur l'ensemble des présentations, soit 77,5 % de notices bibliographiques absentes d'Aleph.

	Échantillon-nage	Nombre total de références	Nombre de notices hors d'Aleph	Nombre de notices dans Aleph	% de notices consultées
Talks at conferences	32	32	31	1	100%
	66	66	56	10	100%
	10	52	6	4	19%
	10	40	6	4	25%
	5	11	3	2	45%
	5	7	1	4	71%
	5	5	0	5	100%
TOTAL	133	213	103	30	62%

Au total, d'après le tableau ci-dessous, l'étude site de l'expérience L3 à couvert 521 notices bibliographiques sur les 940 recensées, soit 55.43% de l'ensemble des notices bibliographiques. Sur cet échantillon de notices, seules 261 ont été entrées dans la base Aleph. Cela représente étrangement juste 50%.

	Echantillon nage total	nombre total de références	Nombre de notices hors d'Aleph	Nombre de notices dans Aleph	% de notices consultées	Taux d'absence général des notices dans la base	Taux de d'absence des notices dans Aleph sans la catégorie publication
TOTAL	521	940	261	260	55.43%	50.1%	25,36%

Il faut noter que ces tables possèdent leurs faiblesses, car le nombre de notices sondées est différent suivant les types de documents. De plus, la masse des publications CERN concernant L3 a été étudiée de façon exhaustive, donc son grand nombre d'échantillons fausse les chiffres globaux en faveur de l'appartenance à la base. Au CERN, lorsqu'il y a publication, l'entrée dans la base de donnée est sensée être obligatoire et quasi-automatique.

Si l'on enlève, des résultats globaux, les chiffres tirés de l'étude des publications, la présence des notices dans la base ne représente plus que 88 notices pour un ensemble de 347 un petit 25,36%. Globalement, cela veut dire que pour les

419 notices bibliographiques non étudiées du site de L3, 74.64% ne font pas partie de la base soit environ 312. On peut estimer alors que le site de L3 possède 573 notices à saisir ou à importer dans Aleph (312 + 261). C'est-à-dire 61 % des références bibliographiques totales du site L3.

Manuellement, à raison d'un quart d'heure pour chaque notice, cela représente pour la bibliothèque un temps horaire global de 143 heures un quart (573 x 0.25). Ainsi, un employé de la bibliothèque à plein temps devra consacrer 18 jours (143 : 8 = 17,9), un mois, à l'importation de documents du Site de L3. Sachant qu'il existe désormais 8 expériences du type et de la taille de L3, il lui faudrait plus les trois quart de l'année pour recueillir les informations disponibles. Informations qui, il faut le souligner, ne sont pas figées et évoluent de façon cumulative.

Le site L3 du CERN est une collaboration internationale impliquant les instituts suivants.

- | | | |
|--|---|--|
| 1. Aachen 1 - RWTH | 19. HEFEI (USTC) | 34. Nijmegen - University |
| 2. Aachen 3 - RWTH | 20. ITEP | 35. NIKHEF |
| 3. University of Alabama | 21. Johns Hopkins University | 36. Northeastern University |
| 4. Basel University | 22. Advanced Inst. of Sciences and Technology (KAIST) | 37. Paul Scherrer Institut |
| 5. Beijing (IHEP) | 23. Kyungpook National University | 38. Perugia - INFN - |
| 6. Humboldt University | 24. Laboratoire d'Annecy-le-Vieux de Physique des Particules (LAPP) | 39. Princeton University |
| 7. Bologna - INFN | 25. Université de Lausanne | 40. Purdue University, USA |
| 8. Boston University | 26. Lawrence L. National Laboratory | 41. Roma - INFN - |
| 9. Institute of Atomic Physics and University of Bucharest | 27. Los Alamos National Laboratory | 42. SAN DIEGO |
| 10. Budapest | 28. Louisiana State University | 43. SEFT |
| 11. California Institute of technology | 29. Université Claude Bernard (IPNL) | 44. Shanghai - Institute of Ceramics - |
| 12. Carnegie-Mellon University | 30. Massachusetts Institute Of Technology | 45. SOFIA - Bulgarian Academy of Sciences |
| 13. CIEMAT | 31. Michigan University | 46. Nuclear Physics Institute of St Petersburg |
| 14. University of Sciences, Nicosia | 32. Università degli study di Milano - INFN | 47. National Central University, Taiwan |
| 15. DESY-Institut fur Hochenergiephysik | 33. Naples - INFN | 48. TATA |
| 16. Eidgenossische Technische Hochschule (E.T.H. Zurich) | | 49. University of Utrecht |
| 17. Università di Firenze - INFN | | 50. World Laboratory, FBLJA Project, Switzerland |
| 18. Geneve - Université - | | |

Ces cinquante instituts possèdent leurs propres publications, des thèses, etc... qui, je le supposais, ne sont pas toujours soumises au site de L3. Donc, la bibliothèque du CERN, qui reçoit du site L3 3 documents sur 5, avait encore moins de chance de les acquérir. Il fallut alors commencer à envisager une extension de la recherche de l'intranet à l'internet.

Comme je l'avais laissé entendre, il semblait tout aussi indispensable de faire une courte analyse des publications du site d'un petit groupe de recherche d'une grande université ou d'un laboratoire lié à l'expérience L3. Ceci afin d'évaluer grossièrement la quantité de notices totales à importer pour une expérience dans les bases de données du CERN et évaluer le temps horaire (voire budgétaire) que cette opération prendrait.

Le site choisi pour cette petite analyse, celui du département de Physique expérimentale de Aachen en Allemagne, ne collabore pas qu'avec l'expérience L3. Ce laboratoire de physique des particules contribue aussi à 3 grandes expériences du CERN : CMS, OPAL, AMS ainsi qu'à de nombreuses autres expériences de grands laboratoires internationaux. Donc, il fallait s'attendre à avoir pour certains sites de petits groupes de recherche, une page de publication commune pour toutes les notices de publications et de pré-publications de toutes les expériences auxquelles ils participent.

Dans le site de l'expérience à l'université de Aachen III, la page des publications possède deux types de documents distincts. Les thèses doctorales et les thèses de diplômés. Pour effectuer cette courte analyse, un sondage sur 5 documents extraits de chacune deux listes paraissait suffisant pour commencer à avoir un bon aperçu général de la situation de L3.

Thèses Doctorales			
Auteur	Titre	Présence au CERN	
		Site L3	Dans Aleph
Wynhoff, S	Messung der Tau-Paarproduktion mit dem L3-Detektor bei LEP.	OUI	NON
Möller, M	Gemeinsame Bestimmung der Vorwärts-Rückwärts-Asymmetrie schwerer Quarks mit L3- am LEP	OUI	NON
Roth, S	Messung der Myonpaarproduktion und ihrer Strahlungskorrekturen mit dem L3-Detektor bei LEP	OUI	NON
Riemann, S	Suche nach dem Z'-Boson auf der Z-Resonanz mit dem L3-Detektor am LEP-Beschleuniger,	NON	NON
Sassowsky, M	Vorbereitungen zur Messung der Reaktion $e^+e^- \rightarrow W+W^- \rightarrow jj \mu$ nu mit dem L3-Detektor bei LEP 200,	NON	NON

Le premier tableau ci dessus nous démontra parfaitement que sur 5 titres, seuls 3 sont connus par le site principal de L3 et aucun par la bibliothèque du CERN.

Quant aux documents, ils sont disponibles en grande majorité en version complète (full texte), au format Postscript et compressé.

Au cours de la recherche des documents dans L3, je me suis aperçu que presque toutes les notes de l'expérience sont absentes de la base Aleph. Or, les notes de certains sites comme L3 sont protégées par un mot de passe. Donc on ne peut ni évaluer précisément le contenu d'un site, ni importer les notices librement sans la permission du webmestre qui ne pourrait les refuser à la bibliothèque.

Recherches

Auteur	Titre	Présence au CERN	
		Site L3	Dans Aleph
Jenkes, K	Studium der Reaktion $e+e- \rightarrow e+ed$ m L3 Detektor bei LEP.	OUI	NON
Opitz, H	Messung des Wirkungsquerschnittes $e+e- \rightarrow$ hadronen im Energiebereich zwischen 130 und 140 GeV mit dem L3-Detektor bei LEP	OUI	NON
Schmidt-Karst, S	Studium der Abstrahlung harter photonen im Anfangszustand der Reaktion $e+e- \rightarrow e+e- (\gamma)$ bei LEP	OUI	NON
Von der Mey, M	Untersuchung des Wirkungsquerschnittes $e+e- \rightarrow \gamma \gamma (\gamma)$ mit dem L3 Detektor bei LEP.	OUI	NON
Straessner, A	Studium der photonabstrahlung in der Reaktion $e+e- \rightarrow \mu+\mu- \gamma$ bei LEP	OUI	NON

Le taux de présence des deux pages de documents du groupe de Aachen III affiche 80% de présence dans le site principal de L3 et un bon 0% dans Aleph. Cependant, Il y a lieu de rester prudent car les autres types de publication ne sont pas présents et le titre, généralement en allemand à pu être traduit en anglais ou simplement reformulé différemment.

Si les informations tirées de Aachen III sont justes, cela signifie que ce site possède 22 notices sur 28 dans le site de L3. Donc, les 50 instituts collaborant avec L3 représentent à eux tous théoriquement 20% de plus que tout ce que centralise le site principal de L3. Ainsi, j'ai pu évaluer le nombre total de notices et de documents plein-texte à importer dans Aleph à 687 (573 + 20%) ; c'est à dire l'équivalent de 172 heures (21 journées et demi) de travail supplémentaire pour la bibliothèque.

Au total, l'information semble ici avoir fonctionné schématiquement comme un entonnoir troué. Au lieu de concentrer vraiment l'information du site de petite expérience au site principal, Il y a eu parfois perte d'information en cours de route. Ceci peut s'expliquer par le fait que les chercheurs de L3 se connaissent et savent ou consulter les publications récentes d'un groupe particulier. Très souvent l'information ne remonte pas jusqu'à la bibliothèque. On peut penser que les spécialistes qui veulent se renseigner sur ce sujet préfèrent aller sur le site de L3 et non à la bibliothèque. La promotion de la bibliothèque, pourtant, n'est pas à faire. Une étude récente effectuée par une étudiante technique, Caroline Christansen, a démontré que les chercheurs sont, après les documentalistes eux-mêmes, les principaux utilisateurs du serveur d'Aleph de la bibliothèque.

Cependant, j'ai conclu qu'un courrier électronique groupé aux webmasters et/ou aux responsables des chercheurs serait utile pour qu'ils puissent rappeler aux chercheurs l'utilité de la soumission de leurs travaux à la bibliothèque

La conclusion de l'évaluation du problème était prévisible. Le CERN est un des plus grands nœuds de connexion internet d'Europe. Se limiter à son intranet est illusoire car les informations sont ouvertes à tous, diffusées largement sur l'extérieur et semblent sortir de l'intranet très facilement. L3, qui travaille avec 50 instituts extérieurs, n'est pas la plus riche des expériences car Atlas travaille avec 150 instituts extérieurs. L3 possède à peu près 1150 documents en ligne (940 + 20%) dont 687 sont à importer. Cela représente le travail d'un documentaliste pour un mois. Restait à évaluer le nombre de site d'expériences CERN.

2.1.3 Recherche des sources de publications sur les expériences CERN.

2.1.3.1 Les sources papier

Afin d'acquérir les sources essentielles permettant de trouver le nom des expériences ainsi que leur numéro CERN, il fallait en premier lieu interroger des personnes-ressources de la bibliothèque. Ces personnes sont les documentalistes du service des prêtirages : Jocelyne Jerdelet et Catherine Cart. Elles m'orientèrent vers

deux sources papier : le rapport annuel en trois volumes, qui possède des informations précises sur les expériences et leurs collaborations et surtout une publication du CERN intitulée *Experiments at CERN* communément appelé *gray-book* (pour sa couleur) qui répertorie toutes les activités des expériences et de leurs collaborations pour l'année en cours. Avec ces seules sources, les noms de toutes les expériences ainsi que celui des instituts collaborant avec eux étaient théoriquement déjà connus.

2.1.3.2 Les sources électroniques

L'idéal aurait été de trouver des sources électroniques, mais à part le listing incomplet d'une soixantaine d'expériences de la page d'entrée du site général du CERN, je n'en disposais pas. Il fallut alors engager temporairement une prospection sur internet et intranet avec des moteurs de recherche puis, face au manque de résultats, aux métamoteurs de seconde génération de type Copernic99. En recherchant sur le site du CERN avec cet outil, j'ai pu trouver noms, sites et listes HTML d'expériences avec leurs laboratoires associés ont pu être trouvés. Ainsi, hormis les métamoteurs, et la liste de la *homepage* du CERN : deux types de sources électroniques ont pu être isolés. Le premier type trouvé est la liste des collaborations d'une expérience que l'on trouve parfois dans un site principal. Le second type est le ou les liens que possède un site d'expérience vers d'autres sites frères avec lesquels il entretient des relations régulières. j'ai pu, de la sorte, évoluer parfois d'un site de collaboration d'expériences CERN à un autre site de la même collaboration CERN.

Tardivement, une dernière source est apparue au fil de la recherche, le *gray-book* sous sa version électronique (URL), qui possède de nombreux liens jusqu'aux pages d'expériences et d'instituts collaborant avec le CERN. Cette source peut être considérée comme double car en remontant dans la source de l'URL (c'est à dire en raccourcissant progressivement l'URL), j'ai pu trouver un dossier différent et plus ancien sur les sites d'expérience CERN et non-CERN qui me fut aussi assez utile.

2.1.4 Analyse du nombre de sites d'expériences.

2.1.4.1 Evaluation du nombre d'expériences

A partir de ces sources, j'ai pu, grossièrement évaluer le nombre total d'expériences CERN qui s'est avéré colossal alors que les pages WEB du CERN annoncent 63 expériences composées de 24 projets de recherche et de 39 expériences.

1. AL-01	44. HELIOS	89. IS340	134. NA9
2. ALEPH	45. ICARUS	90. IS341	135. NA10
3. Alice	46. Isolde	91. IS342	136. NA11
4. Asacusa (AD-3)	47. IS20	92. IS343	137. NA12
5. ATHENA (AD-1)	48. IS80	93. IS344	138. NA13
6. ATLAS	49. IS300	94. IS345	139. NA14
7. ATRAP (AD-2)	50. IS301	95. IS346 (Mistral)	140. NA15
8. OPAL	51. IS302 (Isoltrap)	96. IS347	141. NA16
9. Opal Microvertex	52. IS303	97. IS348	142. NA17
10. CHARM2	53. IS304 (Collaps)	98. IS349	143. NA18
11. CLIC STUDY GROUP	54. IS305	99. IS350	144. NA19
	55. IS306	100. IS351	145. NA20
12. CMS	56. IS307	101. IS352	146. NA21
13. COSMO-LEP	57. IS308	102. IS353	147. NA22/EHS
14. CPLEAR	58. IS309	103. IS354	148. NA23
15. CTF team	59. IS310	104. IS355	149. NA24
16. DELPHI	60. IS311	105. IS356	150. NA25
17. E632 Collaboration	61. IS312	106. IS357	151. NA26
18. E852 Collaboration	62. IS313	107. IS358	152. NA27
19. EMULSION	63. IS314	108. IS359	153. NA28
EXPERIMENTS	64. IS315 (COMPLIS)	109. IS360	154. NA29
20. EMU01	65. IS316	110. IS361	155. NA30
21. EMU02	66. IS317	111. IS362	156. NA31
22. EMU03	67. IS318	112. IS363	157. NA32
23. EMU04	68. IS319	113. IS364	158. NA33
24. EMU05	69. IS320	114. IS365	159. NA34
25. EMU06	70. IS321	115. IS366	160. NA35
26. EMU07	71. IS322	116. IS367	161. NA36
27. EMU08	72. IS323	117. IS368	162. NA37
28. EMU09	73. IS324	118. IS369	163. NA38
29. EMU 10	74. IS325	119. JETSET	164. NA39
30. EMU 11	75. IS326	120. L3	165. NA40
31. EMU 12	76. IS327	121. L3 cosmic	166. NA41
32. EMU 13	77. IS328	122. LEAR-PS185	167. NA42
33. EMU 14	78. IS329	123. LEAR-PS205	168. NA43
34. EMU 15	79. IS330	124. LEAR-PS2	169. NA43.2
35. EMU 16	80. IS331	125. LHCb	170. NA44
36. EMU 17	81. IS332	126. NA1	171. NA45 (CERES)
37. EMU 18	82. IS333	127. NA2	172. NA46
38. EMU 19	83. IS334	128. NA3	173. NA47 (SMC)
39. EMU 20	84. IS335	129. NA4	174. NA48 (CPviolation)
40. FELIX	85. IS336	130. NA5	175. NA49
41. Gajet	86. IS337	131. NA6	176. NA50
42. GAMMS 4000	87. IS338	132. NA7	177. NA51
43. GPMIMD-II	88. IS339	133. NA8	178. NA52

179.	NA53	244.	PS169	309.	R409	374.	SC50
180.	NA54	245.	PS170	310.	R410	375.	SC51
181.	NA55(aka.NYMPH)	246.	PS171	311.	R411	376.	SC52
182.	NA56 (SPY)	247.	PS172	312.	R412	377.	SC53
183.	NA57	248.	PS173	313.	R413	378.	SC54
184.	NA58 (COMPASS)	249.	PS174	314.	R414	379.	SC55
185.	P29	250.	PS175	315.	R415	380.	SC56
186.	P30	251.	PS176	316.	R416	381.	SC57
187.	P31	252.	PS177	317.	R417	382.	SC58
188.	P32	253.	PS178	318.	R418	383.	SC59
189.	P33	254.	PS179	319.	R419	384.	SC60
190.	P34	255.	PS180	320.	R420	385.	SC61
191.	P35	256.	PS181	321.	R421	386.	SC62
192.	P36	257.	PS182	322.	R422	387.	SC63
193.	P37	258.	PS183	323.	RD1	388.	SC64
194.	P38	259.	PS184	324.	RD2 (SITP)	389.	SC65
195.	P39	260.	PS185	325.	RD3 (Accordion)	390.	SC66
196.	P40	261.	PS185/2(lear/Hyperon)	326.	RD4 (LHC-muon)	391.	SC67
197.	P41	262.	PS185/3(lear/Hyperon)	327.	RD5	392.	SC68
198.	P42	263.	PS186	328.	RD6 (TRD-tracker)	393.	SC69
199.	P43	264.	PS187	329.	RD7	394.	SC70
200.	P44	265.	PS188	330.	RD8 (GaAsworks)	395.	SC71
201.	P45	266.	PS189	331.	RD9 (SOI)	396.	SC72
202.	P46	267.	PS190	332.	RD10	397.	SC73
203.	P47	268.	PS191	333.	RD11 (EAST)	398.	SC74
204.	P48	269.	PS192	334.	RD12	399.	SC75
205.	P49	270.	PS193	335.	RD13 (LHCTB)	400.	SC76
206.	PS131	271.	PS194	336.	RD14	401.	SC77
207.	PS132	272.	PS194/3(Ionisation)	337.	RD15	402.	SC78
208.	PS133	273.	PS195 (CP Violation)	338.	RD16 (FERMI)	403.	SC79
209.	PS134	274.	PS196	339.	RD17 (FAROS)	404.	SC80
210.	PS135	275.	PS197 (Crystal Barrel)	340.	RD18 (Chryst. Clear)	405.	SC81
211.	PS136	276.	PS198	341.	RD19 (PIXEL)	406.	SC82
212.	PS137	277.	PS199	342.	RD20	407.	SC83
213.	PS138	278.	PS200	343.	RD21	408.	SC84
214.	PS139	279.	PS201(lear/obelix)	344.	RD22	409.	SC85
215.	PS140	280.	PS202	345.	RD23 (O.electronics)	410.	SC86
216.	PS141	281.	PS203	346.	RD24 (SCI)	411.	SC87
217.	PS142	282.	PS204	347.	RD25	412.	SC88
218.	PS143	283.	PS205	348.	RD26 (CSIRICH)	413.	SC89
219.	PS144	284.	PS206	349.	RD27	414.	SC90
220.	PS145	285.	PS207 (Lex)	350.	RD28(G.microstrips)	415.	SC91
221.	PS146	286.	PS208 (Hotnuclei)	351.	RD29 (DMILL)	416.	SC92
222.	PS147	287.	PS209	352.	RD30 (optical trigger)	417.	SC93
223.	PS148	288.	PS210	353.	RD31 (NEBULAS)	418.	EMU04
224.	PS149	289.	PS211 (TARAC)	354.	RD32 (ALICE TCP)	419.	RE1 (AMS)
225.	PS150	290.	PS212 (DIRAC)	355.	RD33 (TGT)	420.	RE2 (Caprice)
226.	PS151	291.	R107	356.	RD34 (TILECAL)	421.	RE2 (Caprice II)
227.	PS152	292.	R108	357.	RD35 (SICAL)	422.	RE3
228.	PS153	293.	R109	358.	RD36(Shashlik Cl.)	423.	TOSCA
229.	PS154	294.	R110	359.	RD37	424.	LHC
230.	PS155	295.	R207	360.	RD38 (CICERO)	425.	E 835
231.	PS156	296.	R208	361.	RD39 (SMSD)	426.	FAROS
232.	PS157	297.	R209	362.	RD39 (CTD)	427.	KLMM
233.	PS158	298.	R210	363.	RD40 (Q-CAL)	428.	LIAF
234.	PS159	299.	R211	364.	RD41 (MOOSE)	429.	NMC
235.	PS160	300.	R301	365.	RD42	430.	Nicole
236.	PS161	301.	R401	366.	RD43 (BHCAL)	431.	SLD
237.	PS162	302.	R402	367.	RD44 (GEANT 4)	432.	IS01-13
238.	PS163	303.	R403	368.	RD45 (P.O.M.)	433.	LVD
239.	PS164	304.	R404	369.	RD46	434.	MACRO
240.	PS165	305.	R405	370.	RD47 (HEP PC)	435.	NAUSICA
241.	PS166	306.	R406	371.	RD48 (ROSE)	436.	NOE
242.	PS167	307.	R407	372.	RD49	437.	RICH
243.	PS168	308.	R408	373.	RPC	438.	UA1

439. UA2	468. WA21	497. WA50	526. WA78
440. UA3	469. WA22	498. WA51	527. WA79
441. UA4	470. WA23	499. WA52	528. WA80
442. UA5	471. WA24	500. WA53	529. WA81
443. UA6	472. WA25	501. WA54	530. WA82
444. UA7	473. WA26	502. WA55	531. WA83
445. UA8	474. WA27	503. WA56	532. WA84
446. Opera	475. WA28	504. WA57	533. WA85
447. LHCB	476. WA29	505. WA58	534. WA86
448. WA1	477. WA30	506. WA59	535. WA87
449. WA2	478. WA31	507. WA60	536. WA88
450. WA3	479. WA32	508. WA61	537. WA89
451. WA4	480. WA33	509. WA62	538. WA90
452. WA5	481. WA34	510. WA63	539. WA91
453. WA6	482. WA35	511. WA64	540. WA92 (Beatrice)
454. WA7	483. WA36	512. WA65	541. WA93
455. WA8	484. WA37	513. WA66	542. WA94
456. WA9	485. WA38	514. WA67	543. WA95 (CHORUS)
457. WA10	486. WA39	515. WA68	544. WA96 (Nomad)
458. WA11	487. WA40	516. WA69	545. WA97
459. WA12	488. WA41	517. WA70	546. WA98
460. WA13	489. WA42	518. WA71	547. WA99
461. WA14	490. WA43	519. WA72	548. WA99/2
462. WA15	491. WA44	520. WA73	549. WA100
463. WA16	492. WA45	521. WA74	550. WA101
464. WA17	493. WA46	522. WA75	551. WA102
465. WA18	494. WA47	523. WA76	552. DRDC
466. WA19	495. WA48	524. WA76'	
467. WA20	496. WA49	525. WA77	

Toutes les sources ont montré qu'elles possédaient des manques en expériences.

La seule source à être relativement exhaustive est le *gray-book* électronique qui possédait 110 noms d'expériences. Cependant, il lui manque tout de même de nombreux URL de sites d'expériences et de collaborations.

En élaborant cette liste j'ai découvert que les expériences officielles du CERN ne centralisent pas toutes les attentions, il existe aussi de nombreuses petites expériences annexes. Ce sont des expériences de recherche-développement en électronique, optique, comme les expériences RD19 Pixel par exemple, ou encore de petites expériences de collaboration entre de petits laboratoires qui utilisent les infrastructures du CERN comme PS 185 LEAR.

Au cours de l'élaboration de cette liste, je me suis rendu compte que ,parmi celles-ci, certaines n'ont plus cours. D'autre part, il existe des sites d'expériences définitivement achevées possédant encore des données qui n'ont jamais été soumises à la bibliothèque.

C'est pourquoi j'ai choisi de compléter les manques qui m'apparaissaient flagrants. C'est à dire que lorsque dans la liste je trouvais les expériences RD44,

RD45, RD47, RD49 ; je lui rajoutais RD46 et RD48 afin d'effectuer une recherche sur ces expériences dont je supposais l'existence.

2.1.4.2 Evaluation du nombre de sites et du temps horaire pour traiter toutes les notices manuellement

Pour analyser le nombre de sites WEB à consulter, il faut évaluer le nombre moyen de groupes de collaborateurs pour une petite expérience type.

Sur un échantillon d'une vingtaine de petites expériences, j'ai calculé le nombre moyen de laboratoires d'universités qui collaborent avec elles.

Expérience	Nombre de laboratoires	Expérience	Nombre de laboratoires
IS 300	2	RD 19	27
IS 315	7	RD 8	17
IS 343	4	RD 39	13
IS 355	4	RD 47	6
IS 356	4	RD 31	5
IS 344	4	RD 27	11
Moyenne pour le type IS	4,17	Moyenne pour le type RD	13,17

NA45	7	WA95	25
NA48	20	WA8	21
NA52	4	WA96	23
NA49	20	Moyenne pour le type WA	23
NA50	13		
Moyenne pour le type NA	12,80		

Moyenne pour le type AD1	15	Moyenne pour le type PS	13
--------------------------	----	-------------------------	----

collaborateurs par petite expérience	13,52
--------------------------------------	-------

D'après le tableau ci-dessus, j'ai calculé qu'en moyenne, une petite expérience CERN possède 13 instituts qui collaborent avec elle. j'ai effectué à part le total des instituts collaborant avec les huit grandes expériences du CERN.

Total ATLAS	150
Total ALICE	80
Total CMS	150
Total LHCB	47
Total OPAL	34
Total L3	57
Total ALEPH	34
Total DELPHI	57
Total des pages WEB d'instituts collaborant avec une grande expérience CERN.	609

Ainsi j'ai pu évaluer ce total à 609 URL de sites d'instituts collaborateurs dont il fallait trouver la page de publication. Pour l'ensemble des grandes expériences, cela représente un travail de 101,30 heures (6090 minutes) si l'on compte 10 minutes en moyenne pour trouver dans un institut les pages consacrées aux publications de l'expérience cherchée. Une personne à temps plein travaillerait 12.5 jours pour tout découvrir. Comme j'ai calculé d'après Aachen III qu'une page d'institut possède environ 6 notices à récupérer dans Aleph, cela signifie qu'il y a environ 3654 notices à saisir manuellement. Ceci implique un travail de 913 heures 30 ou environ le tiers d'une année de travail à temps plein pour une personne. Si nous cumulons le temps consacré à la recherche d'une page de publication avec celui de la saisie manuelle des documents dans Aleph, nous obtenons, en définitive, 126.5 jours de travail.

D'après le tableau suivant, j'ai évalué à partir de la moyenne le nombre total maximum d'instituts collaborant avec une petite expérience puis, le travail que cela représentait.

Total pour les 61 expériences de type IS (4,17x61)	254
Total pour les 59 expériences de type NA (59x12,80)	755
Total pour les 49 expériences de type RD (49x13,17)	645
Total pour 102 type WA	2346
Total pour les 3 expériences de type AD	45
Total pour les 84 expériences de type PS	1092
Total pour les 160 expériences de type autre	2164
Nombre d'instituts collaborant avec une petite expérience CERN si elles existent toutes.	7301

Nous obtenons 7300 groupes de chercheurs d'instituts travaillant en collaboration avec le CERN. j'ai calculé, qu'un institut collaborant avec une expérience possède en moyenne 20% de publications que le site principal ne recueille pas. A raison d'une moyenne de 6 notices inconnues par sites, cela représente pour les laboratoires associés, un ensemble de 43800 notices inconnues des sites principaux.

A cet ensemble, il faut ajouter les 4686 notices (781x6) des collaborations de grandes expériences CERN ainsi que les notices des sites principaux. Comme, au total, les expériences CERN de la taille de L3 possèdent 3654 notices, cela fait un total de 8340 documents pour les grandes expériences. Pour les ces dernières, il faudrait 260 jours pour tout saisir dans ALEPH et pour les petites expériences, il faudrait 1368 jours pour le même travail.

Au total, cela représente 1628 jours de travail pour une personne. 4 ans et demi de saisie dans Aleph pour une personne à temps plein, cela paraît beaucoup.

En effet, nous verrons plus tard comment ce chiffre théorique diminuera.

Ainsi, l'évaluation globale de ce travail fastidieux, nous obtenons un total de **7910** pages WEB à parcourir pour importer les publications. Ce butinage prendrait $[(7910 \times 10) / 60] / 8$ 164 jours.

Cependant, il faut relativiser ces résultats car ils ont été créés à partir d'un nombre maximum d'expériences. Nombreuses seront celles qui n'existeront pas ou plus. j'ai pu toutefois en conclure que l'importation uniquement manuelle des documents CERN est difficilement applicable. Il faut bien que la bibliothèque trouve un moyen d'automatiser au maximum ces importations de documents.

2.2 Phase de développement

2.2.1 Exploitation des sources par un listing systématisé en fonction des besoins

Afin d'exploiter au maximum les données trouvées sur internet et intranet, j'ai choisi de classer systématiquement les informations tirées de la recherche d'expériences CERN.

J'ai élaboré un tableau excel97 qui se présente ainsi. (C.f. annexe 9)

Nom	Titre	Nom de l'institut	Adresse Web	Adresse du Webmasters	Type de publication	Adresses URL de la publication
ALEPH	Apparatus forLEp PHysics	CERN	Http://alephwww.cern.ch/WWW/	webmaster@alephwww.cern.ch	Theses 1999	http://alephwww.cern.ch/ALEPUB/thesis/thesis.html

La finalité de ce tableau était d'avoir définitivement indexé toutes les URLs de pages WEB afin de trouver ultérieurement un système pour les importer de façon automatique. Le problème est que les adresses de sites WEB changent en principe assez rapidement. Ce tableau devra donc être régulièrement mis à jour. j'ai pu estimer que d'ici 2 ans, ce tableau sera déjà obsolète.

Le recueil des adresses électroniques des webmestres permettra à la bibliothèque de contacter un webmestre particulier ou tous les webmestres d'une expérience par un mail groupé.

Le fait d'avoir le nom complet de l'expérience dans ce tableau permet d'effectuer une recherche Web ultérieure avec le nom complet de l'expérience afin d'obtenir de meilleurs résultats. Par exemple, si je souhaitais rechercher le simple terme Aleph dans le Web, j'obtenais énormément de bruit car le terme Aleph correspond à de nombreuses expériences dans le monde, au système de gestion de la bibliothèque CERN, à des organisations juives internationales, etc... Par contre, avec une recherche à partir du nom complet de l'expérience «Apparatus for LEp Physics», pour peu que le moteur de recherche connaisse la recherche par phrase complète, je trouvais directement la première page de l'expérience Aleph.

Quant au recueil du type de publication, il fera gagner du temps à la personne qui devra saisir tous ces documents dans les bases Aleph. Pour pouvoir compléter les manques dans ce tableau et trouver les URLs d'expériences et d'instituts, il fallait pouvoir utiliser des outils performants et posséder une méthode de recherche sur l'internet pour limiter au maximum le temps d'investigation.

2.2.2 Méthodologie d'une recherche exhaustive sur internet

Prétendre à l'exhaustivité avec internet est une illusion car tant qu'internet ne sera posséder pas de normes et de répertoire général où une inscription est obligatoire : il y aura une perte d'information. Cela ne se fera sans doute jamais, car ce serait contraire à la philosophie de l'outil qui donne l'illusion d'une liberté totale.

Internet possède le paradoxe d'être à la fois l'outil le plus avancé en matière de communication et l'outil au système de classement le plus archaïque. Sans méthode et de bons outils pour l'exploiter : Internet reste l'antithèse de tout bon documentaliste : le classement, la gestion et la recherche d'information précise. Pour s'assurer d'une recherche efficace sur internet, le documentaliste doit comprendre le fonctionnement des outils qu'il utilise. « Savoir, c'est comprendre ! Comprendre c'est améliorer ses recherches. »⁵

⁵ Leloup Catherine, Moteurs d'indexation et de recherche, Paris, 1997

Les trois principaux outils de recherche sur internet sont les annuaires, les moteurs de recherche et les métamoteurs.

2.2.2.1 Les annuaires internet

Ce sont des répertoires de sites indexés en catégories de manière hiérarchique, un peu comme des thésaurus géants. Le plus connu des répertoires est Yahoo.com. Le fonctionnement d'un annuaire est assez simple ; les sites sont soumis à un programme central de l'annuaire qui les indexe. Leurs propres créateurs, ajoutent souvent eux-mêmes un commentaire associé et les classent dans une catégorie propre. Ainsi l'utilisateur peut effectuer une recherche par mots clefs et opérations booléennes ou, comme avec un thésaurus plonge dans la hiérarchie des catégories. Le problème de l'annuaire est que, plus un annuaire est connu, plus il y aura de soumission et plus il sera exhaustif sur une question. Cela explique pourquoi ils ne disposent que d'une information faiblement exhaustive, même si certains se transmettent les données collectées.

2.2.2.2 Les moteurs de recherche

Ils indexent tous les mots de tout le WEB. Ce travail de titan est confié à des robots qui répertorient, classent et indexent le contenu du WEB sans interruption. Il leur faut en moyenne 30 jours pour faire le tour complet des sites avant de recommencer. Lors de son arrivée sur un site trouvé en scannant la liste alphabétique des URL (www.A.com, www.B.com , www.C.com, etc...), le robot indexe toute l'information textuelle d'un site en suivant les liens hypertextes. Lorsque l'on cherche une information, le moteur de recherche renvoie l'URL de toutes les pages collectées concernant ce terme, cette phrase ou ces mots-clefs. L'avantage sur les annuaires est que les pages scannées sont plus nombreuses.

Cependant, il persiste 3 problèmes. Le premier problème est que les informations ne sont indexées en moyenne que tous les 30 jours ; cela ne garanti pas la fraîcheur de l'information reçue. Ensuite, comme le couple vitesse de transmission des informations sur le réseau et puissance des ordinateurs n'augmente pas

proportionnellement à l'accroissement total des informations sur le WEB, les robots mettent de plus en plus de temps à faire le tour complet d'internet. Enfin, le troisième problème est lié au mode d'indexation des robots qui n'indexent que ce qu'ils peuvent lire dans les pages HTML. Les formats textes particuliers ne sont pas lus (les documents PDF et PostScript par exemple) et dans tous les cas les images de textes scannés et les bases de données.

2.2.2.3 Les métamoteurs :

Les métamoteurs sont des agents intelligents dévolus à la recherche sur internet dont la particularité est de pouvoir interroger en une fois différents outils de recherche (annuaires comme moteurs de recherche) afin de fournir la réponse la plus exhaustive à une question donnée. Le problème qui se pose à l'utilisateur est que chaque outil de recherche possède ses propres spécificités. Les propriétaires des outils interrogés apprécient rarement ces outils dans la mesure où ils empêchent la consultation des bandeaux publicitaires de leurs sponsors, leur principale source de revenus. Les métamoteurs actuels sont de deux types les plus largement diffusés. Ils peuvent être disponibles à partir d'un serveur comme la première génération de métamoteur existant, de type *metacrawler*.. Ou encore, ils peuvent être disponibles comme logiciel client à télécharger et à installer sur son propre ordinateur. Pour la plupart, il y a possibilité de régler le temps maximum à passer pour chaque recherche et de choisir les différents moteurs. Le traitement des résultats obtenus va du listing brut à l'élimination des doublons avec classement selon la pertinence supposée de l'information pour les plus évolués.

2.2.2.4 Le choix de l'outil Copernic99

Le choix des métamoteurs s'est rapidement imposé pour la recherche d'expériences CERN sur intranet. En effet, le petit moteur de l'intranet du CERN, clone du moteur de recherche en ligne infoseek, était l'outil utilisé en début de stage pour la recherche d'expériences. Ce petit moteur pouvait créer une recherche avancée à partir des opérateurs booléens classiques AND OR NOT et possédait

l'avantage de pouvoir exécuter une recherche dans les adresses URL ce qui s'avérait très utile. Cependant, les résultats tirés de ces recherches restaient étonnamment faibles par rapport au bruit obtenu. Après la découverte de sites principaux d'expériences CERN non hébergés dans l'intranet, avec l'accord des responsables du CERN, la recherche des expériences CERN put être étendue à l'internet avec des outils plus efficaces. Ce fut entre autre le métamoteur Copernic99.

L'outil Copernic99 : J'ai choisi cet agent car ses recherches sont lancées simultanément sur 32 moteurs de recherche et annuaires, 16 en version gratuite. C'est une des palettes de moteurs et d'annuaires utilisés les plus larges du marché. Non seulement il combine ainsi les avantages des deux types de recherche, mais de plus, rares sont les autres métamoteurs qui affichent les résultats au fur et à mesure qu'il les trouve, les classe par ordre de pertinence, élimine les doublons et les liens périmés, et garde en mémoire la recherche avec une possibilité de mise à jour. Pour chacun des résultats apparaît le titre, une description, l'URL, le nombre d'occurrences, la date de la recherche, le moteur à l'origine de la découverte du site ainsi que l'état de ce dernier (accessible, inaccessible, nouveau moteur...). Une fonction de recherche, de raffinement de l'information par mots-clés et opérations booléennes sur les résultats obtenus est disponible. Pour la consultation hors ligne, il est possible de télécharger tout ou partie des documents trouvés.

En conclusion, Copernic 99 est un des agents récents les plus compétitifs techniquement et commercialement car sa version gratuite est librement téléchargeable dans une dizaine de langues dont le français. Les versions compatibles avec des systèmes autres que les Windows (Macintosh, et Unix) sont à venir.

2.2.3 Analyse des résultats obtenus

Les principales conclusions à tirer au bout d'un mois de butinage sur internet sont, malgré le grand nombre de pages de publications et d'expériences recueillies, qu'il existe sans doute moins de sites d'expériences et de collaborations qu'il n'y paraissait au départ.

D'une part, de nombreux numéros d'expériences que j'ai supposé exister, n'ont jamais été trouvés. Cela a été quand même utile car il fallait en être sûr. De

plus, cela m'a permis au cours de l'exploration des sites, de trouver d'autres expériences CERN.

D'autre part, j'ai observé que l'on retrouve souvent les mêmes laboratoires qui collaborent avec le CERN d'une expérience à l'autre et que certains mettent tous leurs documents dans la même page WEB, ce qui évite de passer deux fois par le même chemin pour deux expériences différentes.

En tout, j'ai relevé 444 URL concernant des documents CERN dans internet et intranet.

Sachant évaluer approximativement le temps que représente une recherche et une indexation exhaustive des URL de tous les documents concernant toutes les expériences, j'ai décidé de concentrer mes efforts sur les collaborations des grandes expériences. C'est-à-dire qu'il me fallait limiter les importations manuelles des documents aux sites principaux de grandes expériences CERN pour ensuite passer aux collaborations de ces dernières.

2.3 Recherche d'une méthode automatisée d'importation des données puis, mise en place d'un système basique de veille documentaire.

Il est ressorti de l'analyse de ces résultats une foule d'idées afin d'importer les données d'internet et d'intranet. Cinq d'entre elles me semblaient applicable afin de contourner le problème de la « non-soumission » des documents des chercheurs à la bibliothèque.

2.3.1 L'importation automatique de publication CERN de l'intranet

Ainsi, pour pouvoir importer automatiquement des publications d'internet jusque dans la base ALEPH (à l'image de ce qui a été fait pour les grandes bases de données INSPEC et UNCOVER) il fallait que les publications des chercheurs sur internet soit normalisées. Or, sur internet, il n'existe aucune norme de publication, et la loi générale semble être celle de la liberté. Cette méthode a été considérée par le support informatique de la bibliothèque comme théoriquement faisable mais dans la pratique inapplicable, car cela nécessitait quasiment la création d'un programme pour

chaque page Web CERN. Comme cette méthode fut vite écartée, il fallut donc, après une redéfinition du projet de ce stage, trouver une solution pour contourner ce problème et normaliser les notices de publications des chercheurs CERN sur intranet et internet .

2.3.2 Une aide à la publication sur internet

Théoriquement, cette méthode consistait à fournir un outil aux webmestres afin d'automatiser et normaliser sur internet la publication des notices bibliographiques en HTML. Les documents une fois normalisés, l'importation automatique à la bibliothèque était possible. Une astuce consistait à envoyer les informations recueillies en parallèle sur une page HTML à la bibliothèque du CERN, comme une petite soumission simplifiée. On pouvait espérer ainsi «rapatrier» les documents qui nécessitent un code pour leurs publications. Comme les notes, documents auxquels la bibliothèque n'a pas accès dans la plupart des sites de grandes expériences et qui sont peu présents dans Aleph.

En récoltant les sites de publication, je me suis rendu compte que certains sites, comme l'expérience WA98, possèdent déjà un système de soumission simplifié de documents pour faciliter aux chercheurs l'édition de leurs notices sur internet.

Il existait donc déjà un modèle préétabli de ce type de programme. Après une recherche plus approfondie, j'ai découvert que ce genre de programme est appelé *Guestbook*. Un *Guestbook* est le nom anglais que l'on donne sur Internet pour "livre d'or". Après avoir visité un site WEB, avec ce type de programme, un visiteur peut laisser son nom, son adresse, son adresse URL, un E-mail, des commentaires, etc...sur une page HTML. Rien ne m'empêchait de remplacer le titre de ces champs par Auteur, Titre, Adresse du document en plein texte etc.

Ainsi, la majorité du programme à élaborer était déjà disponible dans sa forme brute sur internet. Il suffisait juste de créer une seconde page HTML de destination située sur le site de la bibliothèque et commune à toutes les expériences.

2.3.2.1 Les deux sources existantes pour acquérir un programme ressource sur le web.

Deux sources se sont imposées : l'importation d'un programme préétabli et fonctionnel situé sur un site et l'importation d'un des multiples scripts gratuits de *guestbook* disponible sur internet.

2.3.2.1.1 L'importation d'un programme fonctionnant sur internet.

Pour la première des deux sources, il s'avéra assez vite que la simple importation des deux pages nécessaires au fonctionnement de ce programme (page de saisie, et de destination) n'était pas suffisante, car les liens entre ces deux pages ne semblaient pas établis. La solution ponctuelle trouvée face à cette difficulté paraissait être l'utilisation d'un autre type d'agent intelligent : un *offline browser*.

La traduction française d'un *offline browser*, très significative, est « aspirateur de site. » Ce terme qui prête à sourire indique bien que désormais, les internautes n'ont plus besoin de rester des heures connectés sur internet à consulter un site. En effet, ces technologies actuelles permettent « d'aspirer » des sites complets en peu de temps, de reconstituer les liens hypertextes, d'importer des programmes annexes et de les consulter une fois hors ligne. Ne connaissant pas d'emblée l'existence des répertoires CGI-BIN (qui protègent les applications qui fonctionnent en ligne) l'avantage d'un tel outil parût évident pour la constitution de ce programme. Le débat sous-jacent de ce type de méthode est celui du droit légal que dispose sur son œuvre l'auteur de cette production. Sachant que les *guestbooks* sont largement disponibles sur internet, j'ai pu considérer que les sites et les programmes ne comportant pas de licence les protégeant contre l'importation sont libres de tout droits. D'ailleurs, aucun droit international ne s'y applique pour l'instant.

Une importation de script ou de pages HTML représente pour un webmestre un gain de temps considérable. D'ailleurs, nombreux sont ceux d'entre-eux qui piochent dans les sources de leurs collègues afin de gagner du temps pour la programmation spécifique et le contenu réel du site. La limite entre plagiat et inspiration est très difficile à évaluer sur internet. Pour le cas de notre programme, le

guestbook devait servir essentiellement de source et ne devait pas être utilisé tel quel. C'est ce qui me conforta dans l'utilisation de deux « aspirateurs » de site : Memoweb98 et Htrack.

2.3.2.1.2 La deuxième source à notre disposition

Des répertoires de scripts de programmes libres de tout droits disponibles sur internet. Ces répertoires sont, pour en nommer quelques uns , « Best-Of-Web.com », « perlarchive.com », « CGI-City.com », « WebScripts.com »,... Ceux que j'ai cherché à utiliser sont freescrpts.com qui propose des scripts CGI écrits en PERL, « freecode.com » qui possède de nombreux applets Java gratuits ainsi que des scripts PERL, C++ et Visual basic, où encore « worldwidemart.com » qui propose des scripts Perl, C++, Html et Java.

Cependant, avec ces sources, il fallait sélectionner un langage de programmation, comprendre l'utilité et le fonctionnement d'un fichier CGI et obtenir un emplacement WEB pour le faire fonctionner.

2.3.2.2 Qu'est-ce que le CGI-BIN et à quoi sert-il ?

Le Common Gateway Interface (CGI) est une norme utilisée pour lier une application au serveur. La plupart des programmes CGI sont écrits avec le langage script PERL. Les programmes CGI peuvent aussi être des programmes « compilés », écrits en C ou des programmes en Visual Basic. Le répertoire courant dans lequel on peut placer un script est généralement appelé C://home/cgi-bin/.

Un document HTML simple est statique, conforme au format texte et ne change jamais. En revanche, un programme CGI peut rendre une information dynamique en temps réel. Le CGI, permet de mettre une base de donnée sur internet et la rendre interrogeable par n'importe qui , de n'importe où.

En théorie, simplement, un programme CGI doit s'exécuter pour transmettre les informations au moteur de la base de donnée et rapatrier ensuite les résultats sur une page WEB. Les fichiers CGI-BIN sont une ressource illimitée pour un centre de documentation possédant une connexion sur le WEB.

La seule chose à se rappeler est que, quoi que fasse le programme, il ne doit pas être trop long à s'exécuter sinon l'utilisateur risque de perdre patience et d'abandonner la connexion !

2.3.2.3 Pourquoi un programme ne peut être appliqué depuis une page de la bibliothèque du CERN ?

2.3.2.3.1 Le fichier « www » et la limitation des « *login* » de la bibliothèque.

Le CERN met à disposition de tous ses chercheurs et ses employés un fichier appelé *www* dans lequel ces derniers peuvent à leur guise placer leurs documents (html, postscript ou autre) afin de les éditer sur le WEB.

Cela est très facile car à chaque employé CERN correspond un numéro d'identité CERN, un mot de passe d'accès et un *login* particulier. Chaque compte CERN inclut une place mémoire personnalisée dans un des disques durs du serveur depuis n'importe quelle machine, une adresse e-mail, et une place dans l'annuaire téléphonique WEB du CERN. Cependant, il faut relever des différences entre ces comptes d'accès à internet !

En effet, chaque entrée dans l'intranet d'un particulier, appelée *login* possède des privilèges propres.

Il semble normal que des étudiants-stagiaires, par exemple, n'aient pas les mêmes droits d'accès aux ressources informatiques qu'un théoricien chercheur en physique nucléaire. Par contre, on peut se demander pourquoi les comptes des documentalistes de la bibliothèque du CERN sont bridés. De plus, à l'avenir, les documentalistes seront amenés à se spécialiser dans les technologies de l'information, à gagner en autonomie vis à vis des informaticiens, et à utiliser pleinement toutes les ressources informatiques d'un site internet.

D'autant plus que les répertoires CGI sont visiblement indispensables pour faire fonctionner un programme de gestion de base de donnée en réseau.

2.3.2.4 La sélection du langage de programmation

Un programme CGI peut-être à priori écrit dans n'importe quel langage de programmation. On peut citer notamment les programmes C et C++ , JAVA, Fortran, PERL, TCL, les Shells Unix, Visual Basic, les scripts Apple etc.

Le seul problème qui pourrait se poser est qu'il faut connaître parfaitement le système de sa propre machine et celui de la machine distante pour savoir s'il supporte la version du programme. Sinon, comme j'ai pu le constater, c'est une mission plus que difficile à remplir.

Le site du CERN fonctionne majoritairement sous UNIX.

Le langage C, créé il y a 20 ans, est le programme qui a servi à construire ce système d'exploitation. Notre intérêt s'est porté dans un premier temps sur ce langage. Cependant, son langage est très complexe pour un néophyte car C est en fait presque un langage de bas niveau utilisé surtout pour la programmation système.

J'ai ainsi appris que les programmes peuvent être écrits en plusieurs types de langage : évolué, c'est à dire de haut niveau ou en langage de bas niveau ; le *BASIC*, le *COBOL*, le *FORTRAN*, le *PASCAL*, le *LOGO* sont des langages de haut niveau tandis que le langage d'assemblage, et le langage machine sont des langages de bas niveau. Un langage de bas niveau est lié à un microprocesseur donné et est directement compréhensible par la machine.

Notre choix s'est donc reporté sur des logiciels plus simples, souvent d'avantage orientés objet (il faut comprendre boutons ou champs), dans la lignée des *Visual-basics* de *Microsoft*.

En premier lieu, il fallut éliminer d'entrée les lourds langage de programmation spécialisés dans le développement tels que *Fortran*, qui nécessitent d'avoir à disposition un programme afin de les « compiler » (c'est à dire rendre exécutable). De plus, Fortran qui est utilisé depuis 1954, est devenu un peu obsolète malgré ses mises à jour et est surtout utilisé pour les grosses machines destinées aux calculs numériques telles que Big Blue d'IBM pour sa rapidité d'exécution.

Ensuite, JAVA fut le plus intéressant pour ses applications WEB réputées pour leur simplicité, les APPLETS, qui auraient pu aussi faire marcher ce programme. Un des avantages certains du JAVA est que l'on peut facilement

télécharger gratuitement le *JAVA Développement Kit* (disponible à l'URL : http://JAVA.sun.com/products/OV_jdkProduct.html) gratuitement pour programmer, exécuter ou modifier des parties de programmes complets ou des APPLETs.

Un APPLET JAVA est un petit programme situé dans un serveur WEB permettant de rendre une page HTML active grâce aux navigateurs internet comme celui de Netscape qui l'exécutent. En théorie, les APPLETs peuvent accomplir les mêmes tâches que les autres programmes. Le problème du JAVA est que sa vitesse d'exécution est assez lente ceci explique pourquoi il était préférable de l'abandonner.

Il fallut aussi se résigner à abandonner le Visual Basic car ce dernier est et ne fonctionne volontairement pas sur toutes les plates-formes. Il possède l'étrange particularité de mal fonctionner lorsqu'on utilise les programmes de concurrents directs tels que Netscape par exemple.

Les avantages de la programmation en PERL

Le PERL eut notre faveur car il est ce que les informaticiens appellent un langage de très haut niveau, c'est à dire en clair un langage simple et de dernière génération. Il fonctionne à priori sur toutes plate-forme et s'est très largement inspiré du langage C. Il peut construire un programme complet en moins de 6 lignes simples alors qu'avec C il en fallait une centaine pour un programme identique.

De plus, sa syntaxe est moins rigide. Un programmeur peut se permettre de faire une erreur, le programme fonctionnera quand même dans ses grandes lignes. J'ai pu constater aussi qu'il existe de nombreux programmes CGI écrits en PERL à partir desquels je pouvais m'inspirer. Il faut noter cependant que comme le PERL n'est pas un langage compilé, il est plus lent que le C. Mais finalement, après quelques conseils de personnes ressources : PERL s'est distingué dans mon choix pour sa simplicité face aux autres nombreux langages.

En définitive, il faut noter qu'il y a toujours un choix à faire entre deux combinaisons : rapidité d'exécution et script difficile ou simplicité du langage et exécution plus lente. Le PERL semblait être un compromis acceptable.

Autre avantage en faveur de PERL était la mise à disposition gratuite de bons manuels de langage PERL sur Internet. Il existe surtout un répertoire appelé CPAN (Comprehensive PERL Archive Network) hébergé à l'institut Pasteur à l'adresse <ftp://ftp.pasteur.fr/pub/PERL/CPAN> et qui regroupe l'ensemble des ressources

relatives au langage PERL. Il existe aussi un groupe de news sur le langage PERL : fr.comp.lang.PERL, mais les discussions sont un peu ardues.

En conclusion, l'avantage d'un tel programme était bien d'obtenir une récolte de notices normalisées. Cependant, il fallait que ce programme remplisse certaines conditions. La première était qu'il devait être accepté par le webmaster qui pouvait refuser de l'installer sur son site. Ce devait présenter un avantage pour le webmestre afin qu'il ne le rejette pas. Ce pouvait être par exemple le fait que préconçu et gratuit, il aurait représenté un gain de temps pour le webmestre.

La seconde était l'utilisation concrète de ce programme par les chercheurs qui devait y trouver leur avantage aussi (gain de temps et simplicité de la procédure) à moins de les obliger à passer par ce programme pour publier sur leur site d'expérience, ce qui est faisable mais peu démocratique.

2.3.2.5 Poser un système d'Alerte :

2.3.2.5.1 En quoi consiste une Alerte ?

Une Alerte est un système avancé de veille qui permet d'importer, en l'occurrence à la bibliothèque, les données nouvelles ajoutées sur une page Web grâce à un mail ou à un téléchargement sur une page HTML. L'avantage semble évident pour une entreprise privée qui souhaiterait surveiller les activités de ses concurrents commerciaux les plus sérieux. Chaque proposition sur le WEB d'un nouveau produit permettrait à l'utilisateur d'un tel système d'être automatiquement informé des dernières nouveautés proposées par ses concurrents.

Un tel système de veille semblait très utile pour la bibliothèque du CERN avec ses problèmes spécifiques de centralisation des documents d'expérience. Une Alerte sur chaque page web d'expérience permettait à la bibliothèque d'être au courant, dès leur annonce sur internet ou intranet, des derniers travaux des chercheurs. à défaut de trouver un système d'importation automatique de toutes les notices, l'alerte permettait de recevoir les plus récentes publications, toutes les

notices disponibles sur le WEB, même les plus anciennes, ce programme permettrait de recevoir les plus récentes publications.

2.3.2.5.2 Comment fonctionne ce programme ?

Ce concept de veille inédit a été développé spécialement d'après une demande spécifique de la bibliothèque par Ludovic Noël Chauvin, étudiant à l'école d'Informatique et Réseaux Industriels de Dijon et stagiaire au service du support informatique. Ce programme a été conçu suite à l'étude sur le nombre de sites d'expériences CERN. Les données récentes arrivées dans les pages WEB indexées sont importées dans un premier temps. Ecrit en C, il fonctionnera au CERN une fois par mois. Entre deux dates de lancement, il compare la taille d'un document HTML par soustraction. (C.f. annexe 10)

Cette application peut identifier trois types d'URL : Les URLs qui ne répondent pas c'est à dire dont le serveur ne fonctionne plus, Les URLs qui ont évolué et Les URLs qui n'existent plus mais dont le serveur fonctionne. Une fois ces URL identifiés, il envoie à la bibliothèque du CERN trois messages correspondant à cette identification. C'est à la bibliothèque, de vérifier ces URLs par la suite.

En soustrayant la page HTML ancienne à la page HTML nouvelle, il peut localiser les informations et les envoyer par mail à la bibliothèque ou sur une page HTML. L'alerte ne récupère pas les pages dont le changement est inférieur à 100 caractères, soit une taille de 100 octets. Cette taille minimale est celle que j'ai évalué pour une notice utile à la bibliothèque

2.3.2.5.3 Un gain inestimable pour le CERN

J'ai évalué que le nombre de pages WEB sur lesquelles il faudrait poser l'alerte est d'environ **7900**. Si un chercheur en moyenne écrit un article par mois, cela représente en théorie 94920 notices bibliographiques qui serait importée en une année à la bibliothèque. Ce programme d'Alerte représente le travail de 164 jours de consultation de pages internet, soit le travail de 5 mois et demi et l'assurance de ne manquer d'aucune information récente.

2.3.2.5.4 Quelles sont les limites de cette Alerte ?

En premier lieu, l'alerte ne fonctionne que sur le format HTML et il ne reconnaît pas les autres formes de fichiers PS, PDF, XML De plus, il ne peut faire la distinguer sur une page HTML une notice bibliographique d'un simple décor ajouté. Ce programme ne peut pas, dans sa forme brute, interroger une mini-base de donnée comme il en existe parfois sur les publications d'une expérience CERN. Il ne peut pas non plus interroger des documents en transitant par le protocole FTP.

Avec toute ces limitations, le programme d'Alerte ne fonctionne que sur 384 URL sur un total de 551 recueillies pour l'instant. j'ai pu ainsi évaluer que sur un total de 7900 URLs, la forme brute de ce programme ne pourra faire une veille que sur 70% de tout ce qui est disponible en ligne sur le WEB du CERN.

Ce programme doit encore être développé et raffiné par le support informatique de la bibliothèque pour être entièrement viable et fonctionnel.

2.3.2.6 Faire la promotion aux webmestres et aux chercheurs de la soumission CERN existante

Au cours d'une recherche, je me suis aperçu que l'expérience WA98 utilise un petit programme de soumission afin que ses chercheurs éditent leurs publications sur internet. Après avoir envoyé un mail pour lui demander de faire un lien jusqu'à la soumission officielle du CERN, nous nous sommes aperçus qu'il existait pour ce cas un grave problème de communication. (cf. annexe 6)

En effet, ce webmestre ne savait pas que son groupe de chercheurs pouvait soumettre leurs publications au CERN ! C'est pourquoi, il s'engagea à installer sur son site, un lien jusqu'à la soumission officielle CERN. (Cf. annexe 6)

Face à ce problème j'ai proposé de créer un courrier électronique groupé à tous les webmestres pour faire la promotion de la soumission CERN.

2.4 Phase d'aboutissement

Sur ces 5 solutions trouvées, seules quatre furent sélectionnées.

2.4.1 Résultats obtenus lors de la phase de développement

2.4.1.1 Résultats du courrier électronique aux webmestres

Au fur et à mesure du recensement des URL d'expériences CERN, j'envoyais un mail groupé pour faire la promotion de la soumission CERN. D'après le contenu de ce mail, j'attendais une réponse de la part des webmestres. (Cf. annexe 3) Le groupe OPAL reçut un mail groupé et malgré la centaine de destinataires, je n'avais obtenu aucune réponse en retour. Pour résoudre ce problème, j'ai cherché à savoir pourquoi les webmestres ne répondaient pas et de conclure qu'il s'agissait sans doute du fait que ce mail s'adresse à tous les webmasters et responsables d'expériences CERN. Les destinataires peuvent voir très rapidement que ce mail ne leur est pas destiné de façon individuelle en observant la liste des destinataires.

Donc, la solution semblait être que l'envoi de ce courrier électronique soit ou paraisse individuel. Je me suis donc mis à la quête d'une personne-ressource qui pourrait me renseigner.

2.4.1.2 Résultats d'un mail groupé avec la fonction BCC d'internet Messenger.

Sachant qu'internet Messenger possédait une fonction pour cacher les mails, je me suis adressé à madame

laire Button qui eut enseigné l'utilisation de ce logiciel au CERN. Celle-ci m'expliqua qu'*internet explorer* possède effectivement la fonction BCC qui permet de cacher les différents destinataires au récepteur du courrier. j'ai utilisé cette fonction pour l'envoi du mail groupé aux webmasters du groupe Opal.

Après utilisation de ce mode d'envoi, le résultat fut malheureusement identique : je ne reçût aucune réponse en retour.

Après mûres réflexions j'ai remarqué qu'un destinataire possède le moyen de se rendre compte que ce mail est collectif, justement par l'abréviation BCC qui apparaît en entête de la lettre.

2.4.1.3 Résultat du petit programme d'envoi individuel de mails

Pour détourner ce problème, je me suis adressé à une autre personne-ressource, Ludovic Chauvin, avec lequel je construisais le système d'alerte. Je lui ai demandé de me construire un petit programme pour envoyer un mail de façon individuelle à plusieurs correspondants.

Ce programme ayant été assez rapidement construit en langage C sous Unix j'ai envoyé par ce moyen un mail individuel aux 150 instituts de l'expérience ATLAS (cf. Atlas). Le script du programme, qui nécessitait un compilateur était celui-ci.

```
#include <stdio.h>
#include <stdlib.h>
main(){
  char commande_line[200];
  char f_destin[200];
  FILE *fp;
  fp = fopen ("nom_fichier", "r");
  if(fp){
    while(!fscanf(fp, "%s", &f_destin) == EOF) {
      printf ("\nMail envoye a: %s\n", f_destin);
      strcpy (commande_line, "mailfile -to ");
      strcat (commande_line, f_destin);
      strcat (commande_line, " -s sujet nom_fichier");
      system (commande_line);
    }
    fclose (fp);
  }
  else{
    printf("\nImpossible de trouver le fichier des destinataires !!!\n");
  }
}
```

J'avais déjà trouvé moi même un petit programme identique dans les bibliothèques de scripts gratuits disponibles sur internet. Cependant, je ne savais pas qu'il fallait un compilateur pour créer un «exécutable» avec le langage C. Compilateur que de toute façon, la bibliothèque du CERN ne pouvait pas obtenir.

Le résultat final fut une fois de plus le même car aucun webmestre ne répondit. Une dernière alternative était la reformulation plus incisive du message envoyé mais mes responsables ont préféré l'envoi de ce message tel quel.

2.4.1.4 Le développement de l'importation semi-automatique

Pour développer ce programme, il fallait d'abord faire face à quelques obstacles.

2.4.1.4.1 Trouver un éditeur HTML gratuit pour l'interface WEB

Il y a plusieurs moyens de créer une page HTML. D'une part, la création d'une page HTML est possible en rédigeant le script directement dans un traitement de texte basic de type *Wordpad* ou *Bloc-notes* (sous Windows) au format .TXT. Pour obtenir une page HTML, il faut changer l'extension MS-DOS du fichier en .HTML. Il suffit ensuite de lancer son navigateur habituel et d'ouvrir le fichier désiré pour voir le résultat obtenu. Le problème de cette méthode est qu'il faut connaître parfaitement le langage HTML. Langage qui, même s'il est très simple et possède de bon livre d'apprentissage, est quand même long à apprendre dans ses détails.

D'autre part, Il y a les éditeurs HTML et j'ai pu nommer sans être exhaustif : « AOL Press », « Claris Home Page », « Dreamweaver » de Macromedia, « Front Page » de Microsoft, « Incontext Spider », « Netscape Gold », « Netscape Composer », « PageMill d'Adobe », « Word IA »(extension de Word qui permet de transformer Word en un éditeur).... j'ai choisi AOL Press parce qu'il est gratuit et est aussi bon, sinon meilleur qu'un payant.

Grâce à ces outils, je pus développer les trois pages d'entrées selon le type de publication à saisir et mes deux pages de destination (cf. annexe 9)

2.4.1.4.2 Comment trouver un emplacement WEB possédant un CGI-BIN gratuit.

Comme nous l'avons vu préalablement, il est impossible depuis la bibliothèque de disposer d'un CGI-BIN car les accès à intranet sont bridés. Je ne m'en suis rendu compte que progressivement car, je cherchais à trouver des solutions par moi-même. Après avoir compris ce qu'est un emplacement CGI-BIN, trouver une offre de site gratuit disposant d'un fichier CGI était aisée puisque dans internet, les offres de ce type de site gratuit abondent. Il suffisait d'effectuer pour cela une recherche booléenne avec les mots clefs suivants «Free» and «CGI» and «WEB» and «space» or «place» or «accumpt» dans un moteur de recherche quelconque.

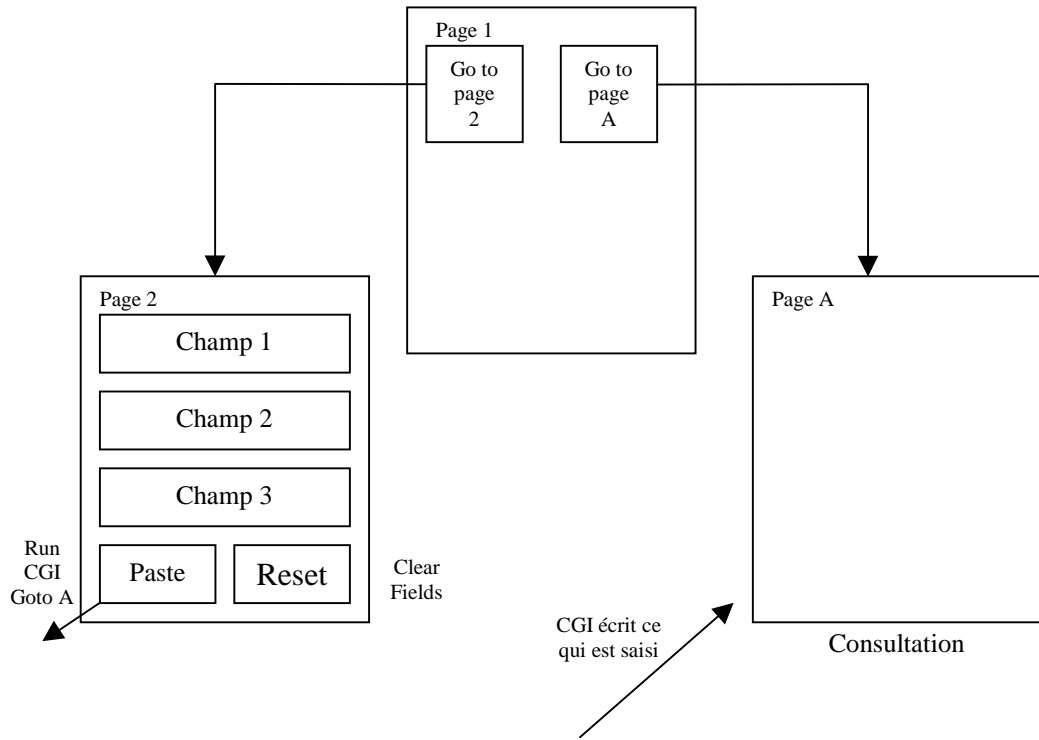
Le véritable problème était le manque d'information de base sur le système d'exploitation du serveur qui devait être compatible avec le langage Perl. Ce langage accepte théoriquement tout type de plate-formes mais, en pratique, il ne fonctionne pas partout.

2.4.2 Comment envoyer les pages de destinations dans les différents répertoires du site WEB distant

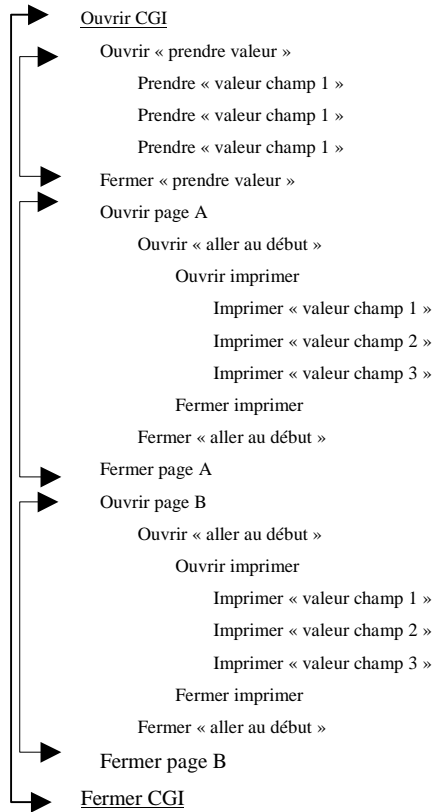
Grâce aux coordonnées fournies avec l'ouverture de ce site, j'ai pu envoyer nos pages WEB sur mon site distant avec les programmes FTP dont je disposais. J'ai pu aussi utiliser un emplacement WWW au CERN afin de disposer pour mon programme d'une page HTML de destination située au CERN.

Le concept de ce programme était sensé être très simple. L'utilisateur devait saisir son texte dans une interface HTML et un bouton devait envoyer les données dans un programme CGI. Le texte, traité par le programme CGI devait être imprimé en HTML dans la page de destination.

Je disposais ainsi de tous les éléments pour pouvoir créer mon programme dont l'algorithme plus détaillé était le suivant :



L'organigramme du programme CGI aurait ressemblé à ceci.



Les principales commandes PERL et CGI ont été trouvées à partir du fichier MAN CGI sous UNIX et du livre de monsieur R.L Schwartz⁶. Le début de ce programme se présentait ainsi :

```
#!/usr/local/bin/perl -w
use CGI ':standard', ':html3';
print header,start_html("passed"), h1('passed args:'),
      "<p>experiment : ", param('experiment'),
      "<p>FirstName : ", param('FirstName'),
      "<p>LastName : ", param('LastName'),
      "<p>Date : ", param('Day'),('Months'),('Year'),
      "<p>Pages : ", param('Pages'),
      "<p>Institut : ", param('Univerity'),
      "<p>Fulltext : ", param('full text'),
      "<p>-----"
-----" ,
      "<p>end" ;
```

Cette commande annonce le début du travail en PERL

Cette commande annonce le début d'un code travail en html.à partir du CGI-BIN

1

2

3

Flèche 1 : cette phrase commande une impression depuis le haut de la page HTML, page dont le titre est h1 ('passed args').

Flèche 2 : La commande <p> imprime le texte entre guillemet sur la page de destination.

Flèche 3 : La commande param ('texte') affiche le texte de ce qui est entré par l'utilisateur dans les champs de saisie de la page de saisie.

Un trait devait être tracé à la fin de la notice pour la différencier des autres. (cf annexe 8)

⁶ SCHWARTZ R. L., Learning PERL., Sébastopol, O'Reilly, 1997.

2.4.2.1.1 Obtenir une aide du support informatique.

Au stade où en était le programme, il pouvait imprimer un résultat sur une des deux pages de destination. (cf.: annexe 8) Cependant, un problème subsistait. Les nouvelles données saisies par l'utilisateur étaient, certes, imprimées une des deux pages internet de destination mais elles effaçaient les anciennes. Cela signifiait deux choses. La première supposition était qu'il existait un script du langage PERL que je ne connaissais pas. Ce script devait permettre de conserver les anciennes données sur les pages de destination. La deuxième supposition était que l'algorithme possédait un manque ou qu'il fallait deux CGI pour faire fonctionner ce programme.

Ne pouvant plus continuer seul ... il me fallait être aidé par le support informatique afin de trouver une solution. Après un entretien avec le responsable du support informatique DH, il m'a fallu abandonner ce projet de programme dans les plus brefs délais. En effet, le CERN disposant d'une soumission électronique de documents en ligne très fouillée créée par le support informatique DH. Pour cela, les responsables de cette section refusèrent de me fournir l'aide nécessaire pour à la finition de ce programme. De leur point de vue, une soumission plus simple que celle existante, diffusée sur tous les sites d'expériences et de collaborations ne représentait aucun intérêt pour le CERN, même si elle permettait clairement l'importation avec un outil de téléchargement automatique des notices normalisées du WEB dans Alice.

2.4.2.2 L'aboutissement de l'Alerte

L'Alerte a été posée en définitive sur 383 pages WEB de publication à causes, nous l'avons vu des différentes spécificités des URL.

2.4.2.2.1 Les premiers résultats

L'expérience de la mise en place de ce petit système de veille scientifique sur internet et intranet faillit être annulé.(C.f. annexe 11)Après une initialisation de la liste que fournie au support informatique et une courte interrogation de celle-ci, une

vingtaine de réponses ont été obtenues. Ayant fait une courte synthèse, les responsables du support informatique ont assez rapidement tiré la conclusion que sa mise en application quelques mois par an était irréaliste car représentait sur un an, un total d'environ 7300 pages à consulter. (Cf. annexes 11)

Après une analyse plus fouillée du premier résultat de cette alerte, j'en ai conclu que ce programme possédait des problèmes systématiquement avec les bases de données , les protocoles FTP, etc.... tous les formats différents du HTML. Après avoir alerté le support informatique que les résultats étaient faussés par ces problèmes de configuration, j'ai proposé de faire tourner ce programme tous les jours et de faire, après une semaine ou deux de configuration du programme et d'assainissement de la liste d'URL, une analyse des résultats.

2.4.2.2.2 Analyse des premiers résultats.

Ainsi, j'ai classé quotidiennement dans un tableau Excel les trois sortes d'URLs extraites par ce programme, basic à l'origine.

Les URLs qui n'existent plus

Les URLs qui ne répondent plus

Les URLs qui ont été modifiés

Au bout de 11 jours de prospection et en établissant un graphique, j'ai pu tirer les conclusions suivantes. Après, l'élimination des URLs qui ne fonctionnaient pas pour diverses raisons ; les « URLs hors services » correspondaient la plupart du temps à celles qui réapparaissent le lendemain dans la catégorie « Nouvelles URLs ». Cette découverte, permit d'affiner la détection des URLs disparues définitivement en établissant une soustraction comme dans le tableau suivant.

jour	Nombre de HS (mais protégées par un code pour le 16, 17, ...ensuite, élimination des URL à code)	Nombre de HS réapparus le lendemain dans la catégorie "changement"	Nombre de HS moins le nombre de ceux qui réapparaissent le lendemain.
July 14, 1999	13	0	-13
July 15, 1999	15	1	-14
July 16, 1999	0	0	0
July 17, 1999	0	0	0
July 18, 1999	28	27	-1
July 19, 1999	5	5	0
July 20, 1999	1	1	0
July 21, 1999	1	1	0
July 22, 1999	1	1	0

Au total, en 10 jours, j'ai pu faire un calcul simple pour pouvoir évaluer la fréquence d'apparition des changements dans les notices. Ainsi, j'ai évalué le nombre de changements journalier. Ils représentent en moyenne plus ou moins 13 changements par jours . Cependant, il faut noter que dans ce tableau, il y a un relevé qui vient changer les données globales.

Il serait donc plus précis pour effectuer une moyenne, d'enlever les écarts les plus importants par rapport à notre moyenne générale, soit les 2 données les plus extrêmes donc ici, les relevés du 14 juillet et du 19 juillet.

Jour	Changements totaux	Changement dans une URL jamais apparue	Changement dans une URL déjà apparue
July 14, 1999	7	7	0
July 15, 1999	9	8	1
July 16, 1999	16	12	4
July 17, 1999	12	4	8
July 18, 1999	11	7	4
July 19, 1999	38	35	3
July 20, 1999	11	5	6
July 21, 1999	8	4	4
July 22, 1999	8	2	5
July 23, 1999	13	4	8
Moyenne	13.3	8.8	4.3
Moyenne moins les écarts les + importants	11	5,75	5

Ainsi, en 10 jours, grâce à l'alerte, j'ai pu évaluer que sur 383 pages WEB de publications, Il y a un changement dans 11 pages tous les jours mais dans ces 11 pages, 50 % ajoutent des publications régulièrement et 50% plus occasionnellement. En gros, si l'on compte 1 nouvelle publication minimum par page, cela fait 11 publications, soit plus de 4015 notices par an soit un travail. Au total, cela représente un travail de 83 jours de saisie des notices dans ces pages

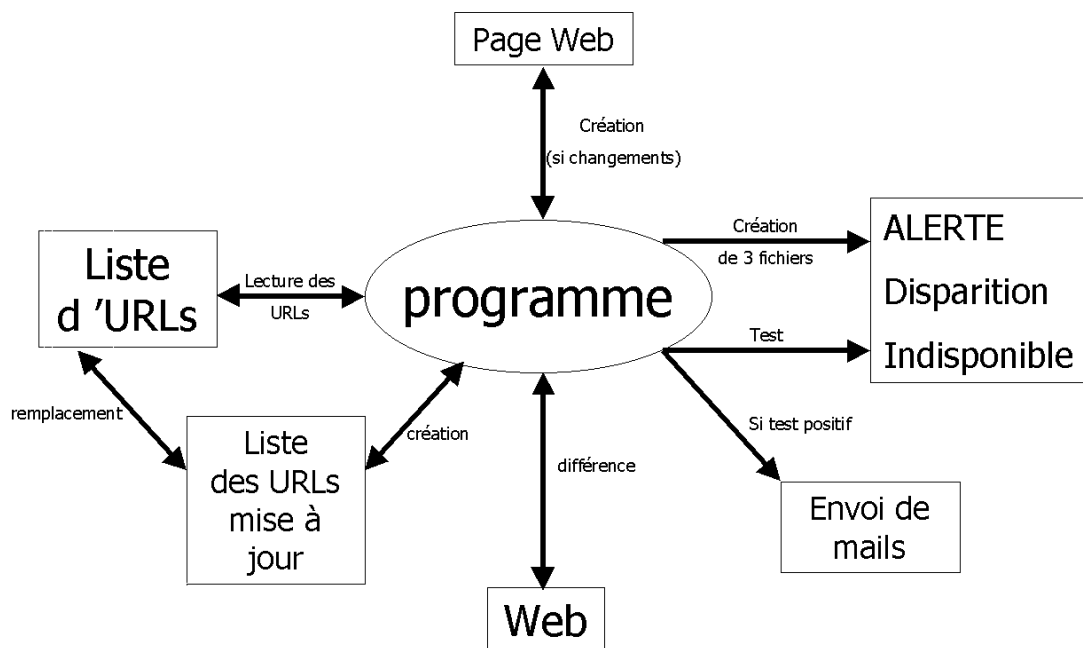
2.4.2.2.3 Développement final

Ce programme d'alerte n'avait seulement qu'un caractère de surveillance. Il lui manquait l'aspect actif sur la page WEB. j'ai alors demandé à Ludovic Chauvin si, au lieu d'un mail simple, il ne pouvait pas, puisque son programme fonctionne sur une soustraction de deux pages WEB d'un jour sur l'autre, nous récupérer les informations et je les envoyer à la bibliothèque en même temps que ses trois messages d'alerte. Cela semblait possible.

Ce programme connut alors trois évolutions. De simple message électronique, il devint message avec données, puis message et page HTML. En effet, après la modification du programme pour récupérer les données issues de la soustraction, le concepteur du programme s'est aperçu que celles-ci étaient la transcription texte d'une page HTML. Ainsi, suite à une troisième modification de ce programme, les données importées soustraites entre les pages WEB récentes et les vieilles ont pu être récupérées et recomposées sur une page WEB de téléchargement à la bibliothèque.

Les documentalistes n'ont ainsi pas eu à perdre de temps à aller consulter toutes Les URLs car le téléchargement récupère les données à leur place et, mieux, recompose les liens jusqu'au texte complet. L'algorithme final de ce programme de 8

pages de script C (annexe 10) est le suivant :



Le programme situé dans les disques durs du serveur du CERN récupère Les URLs stockés dans une liste d'URL extérieure que la bibliothèque peut mettre à jour. A partir de là, il compare la taille des pages HTML initialisées lors du dernier lancement au pages actuelles. Si le programme note une différence entre une page HTML actuelle et une page passée (test), il imprime dans un mail un des trois message concernant les trois situations possibles, puis il l'envoie à la bibliothèque.

Enfin, quand Il y a un changement constaté, il crée une page web pour exporter les données récupérées. (C.f. annexe 12)

L'alerte installée, la bibliothèque du CERN recherche une solution automatique pour éliminer le bruit sur la page Html de destination. C'est à dire pour trouver une méthode permettant de distinguer les notices bibliographiques du texte inintéressant.

2.4.2.3 Récupération semi-automatique des informations dans Aleph

Le problème commun à toutes les méthodes d'importation trouvées était qu'elles recueillaient seulement les notices nouvellement entrées.

Les productions CERN installées depuis longtemps par les chercheurs dans l'intranet du CERN et dans internet ne possédant toujours pas de moyen d'être importées de façon automatique, il fallut se résoudre à le faire manuellement. Cependant, au lieu de rentrer toutes ces publications par la longue procédure d'Aleph, la bibliothèque du CERN trouva une solution semi-automatique.

Il s'agissait d'accomplir le travail que ne pouvaient effectuer les robots, c'est à dire le repérage des publications dans le texte. Mais rien n'empêchait de formater les données recueillies afin de créer des repères qu'un programme reconnaîtra. Ces repères, ce sont les notices minimales créées à partir des champs de la norme anglo-américaine AACR2 . J'ai du mettre au bon format les thèses, les conférences, les papiers et travaux CERN d'une expérience telle qu'L3 pour faciliter l'importation.

Le formatage d'une notice minimale doit se présenter comme ci dessous.

Pour une thèse. (Cf. annexe 7)

EX L3

TI Test des prototypes CAMAC

AU2 Emch, R

IM 1999

Le programme d'importation de ces données formatées dans la base Aleph, développé par Martin Vesely, étudiant en informatique de l'université de Prague, aurait théoriquement pu se mettre bout à bout avec mon programme d'aide a la publication sur internet afin de reconstituer une notice complète.

3. Les autres activités lors de ce stage au CERN.

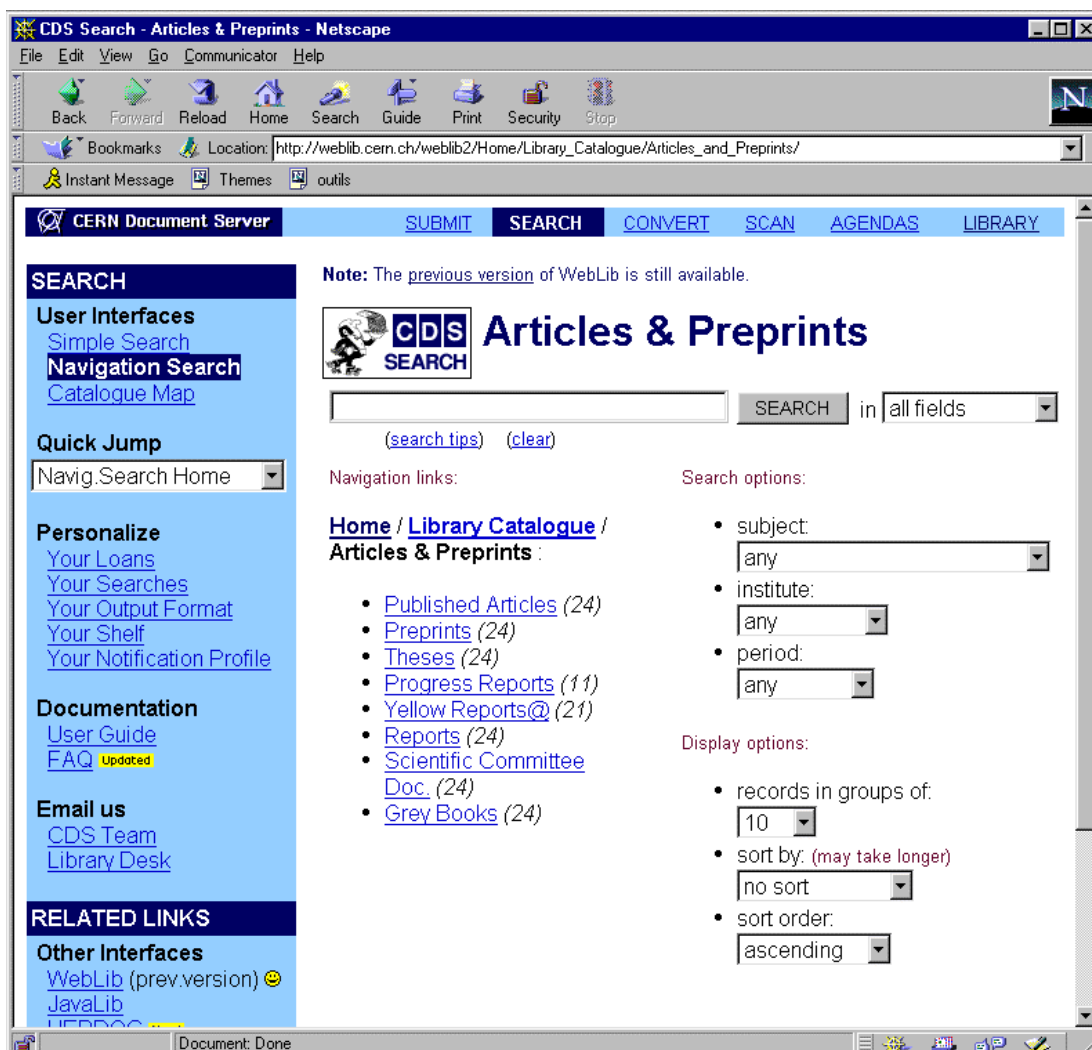
3.1 L'utilisation pratique d'Aleph :

Toute utilisation professionnelle d'Aleph débute par une identification par mot de passe. Un « login ».

3.1.1 La recherche d'une notice bibliographique peut être effectuée selon trois modes différents.

Une recherche sous Aleph 300 peut s'effectuer depuis 2 interfaces : l'interface WEB (ci dessous), pour les utilisateurs et l'interface UNIX en utilisant le protocole TELNET⁷ pour les documentalistes. Une session Telnet permet à l'utilisateur d'entrer en communication avec les bases de données du CERN, le service d'information, les systèmes de gestion, etc. ...

Depuis l'interface utilisateur, il n'y a qu'un mode d'interrogation mais il peut effectuer une recherche de document dans tous les champs, dans le titre, dans le résumé, par auteur et par numéro CERN.



⁷ TELNET = ancien protocole le plus utilisé dans Internet pour communiquer avec d'autres ordinateurs. Actuellement, il permet d'entrer en communication avec des bases de données, des services d'information, des systèmes de gestion, etc.

Les documentalistes du CERN utilisent le mode d'interrogation avancé permettant d'utiliser les trois modes d'interrogation et de combiner les champs de recherche en utilisant les opérations booléennes. Ce mode d'interrogation n'existe pas en ligne pour les utilisateurs.

Derrière l'interface utilisateur se cache une interrogation d'Aleph selon une interrogation directe. La recherche directe est utilisée, pour atteindre un document précis en nommant directement le contenu d'un champ concerné. La liste des exemplaires s'affiche avec leur date de retour, s'ils ont été sortis. (Par exemple, pour trouver un report number (RN) précis, il faudra taper : «*F CERN EP 99 104*»

La recherche par index est la fonction Browse (B). Pour trouver un auteur, il faudra saisir «*B AU= Charpak, G*»

The screenshot shows a terminal window titled 'TNVT - RSPLUS, to host rsplus.cern.ch'. The window has a menu bar with 'Session', 'Edit', 'Commands', 'Settings', and 'Help'. Below the menu bar, there are status fields: 'SEARCH', 'SCAN', 'Format=01', and 'Base= CER'. The main content area displays a list of search results under the heading 'List of Author'. The results are as follows:

Line	List of Author
L1	1 Charos, A
L2	50+ Charpak, G
L3	Charpak, G; Fitch, V; Winter, K; 'Thoot, G; Astbury, A; Kerman, A; Rubbia, C; Maiani, L
L4	10 Charpak, Georges
L5	2 Charpank, G
L6	1 Charpentier, E
L7	1 Charpentier, H
L8	50+ Charpentier, P
L9	1 Charpentier, Ph
L10	1 Charpié, Jean
L11	2 Charpinet, S
L12	1 Charra, J
L13	2 Charra, P

Below the list, there is a command prompt with several options:

```

Continue> C Show documents > line no. Other options > HELP
Reverse > R Show references (*)> XP line no. List of sets > LS
Print > PRI Save as set > F Lno. Start NEW search> START
>>> b AU=Charpak, G
  
```

The time '10:25:27' is displayed in the bottom right corner of the terminal window.

La recherche par mot, c'est la fonction Find (F). Pour trouver un titre précis dont on est parfaitement sûr, il faudra par exemple interroger la base ainsi : «*F Neutrino detection at LEP/TI* »

Les étapes de recherche peuvent être croisées, ce qui revient à construire une équation booléenne. Dans la recherche par mot, il est possible de consulter

l'historique de sa recherche. (Cf.: annexe 9, exemple de recherche avec Aleph). Les troncatures sont implicites ou parfois symbolisées par un point d'interrogation

3.1.2 Le catalogage :

La notice peut être importée de l'extérieur (Los Alamos ...) ou saisie directement dans le catalogue ou être aussi directement créée.

L'indexation des champs se fait à partir d'un des fichiers e fait à partir d'une des bases de données d'autorité qui comprend les renvois "voir aussi" et les "termes associés", "spécifiques" ou "génériques" appelés "*Cross-reference*" en Anglais.

Avec Aleph, il y a possibilité de créer autant de listes d'autorité qu'on le souhaite, cependant, il n'existe pas de véritable thésaurus car les index proposés sont uniquement alphabétiques. Tous les termes employés y sont indexés.

3.1.3 Création de renvois sous aleph 300. (Cf. annexe 4)

Les renvois se traduisent en Anglais par Cross-reference. C'est un outil très utile pour un documentaliste car il permet de faire une liaison entre deux données entrées dans la base. Le rôle des renvois est de standardiser les mots et nom entrés dans la base.

La bibliothèque du CERN utilise ces renvois surtout pour les manques d'accents et pour le traitement des noms d'auteurs russes. En effet, les noms russes possèdent de nombreuses translitérations suivant les pays. Les documentalistes de la bibliothèque utilisent la translitération américaine, cependant, ils ont contacté par courrier électronique les 200 auteurs russes de la base pour leur demander quelle était la forme orthographique à suivre pour leur nom. Grâce à cela, lors de l'interrogation de la base Aleph, un nom russe y apparaît une forme unique au lieu de deux ou trois.

Les renvois sont aussi utilisés pour les titres de périodiques (champ TI) et pour les sujets (champ SU). Une simple abréviation permet de faire référence à un titre ou un sujet complet. Cela évite beaucoup de saisie inutile mais suppose d'avoir un index sous les yeux de façon permanente (comme les stagiaires) ou de connaître par coeur tous ces codes (comme les documentalistes avertis).

Par exemple, simplement

3.2 Apprentissage d'Unix, du Latex, et d'Emacs

3.2.1 Unix

3.2.1.1 Présentation

Unix est un système d'exploitation au même titre que Windows et l'on peut dire sans trop s'avancer qu'il est le seul à vraiment concurrencer Microsoft actuellement.

UNIX a été créé au Laboratoire BELL, aux USA, en 1969. Ses utilisateurs sont surtout des universitaires et des scientifiques. Il existe de nombreuses versions différentes d'Unix. Le grand public connaît de plus en plus LINUX pour les machines PC qui bénéficie d'un phénomène de mode. Mais il a existé, ULTRIX sur DIGITAL, SPIX sur BULL, ... Unix possède la particularité d'être, à la différence de Windows, un système d'exploitation totalement gratuit et téléchargeable depuis internet. On parle souvent de philosophie du monde Unix, qui propose le libre accès de tous aux programmes développés par et pour une collectivité d'utilisateurs. Donc pas de droits d'auteurs sur un programme que tout le monde est libre de modifier à sa guise. Le principe d'Unix est que tous les programmes doivent être ouverts à tous et peuvent être modifiables par tout le monde car la collectivité des utilisateurs retiendra en définitive les meilleures évolutions d'Unix .

3.2.1.2 Les 4 fonctions essentielles du système Unix

Comme tout système : Unix possède 4 fonctions principales.

→ D'abord la gestion des ressources de l'ordinateur car c'est un système d'exploitation qui possède la particularité d'être multi-tâche et multi-utilisateur. Multi-tâche car il peut faire fonctionner simultanément plusieurs programmes, et multi-utilisateur car plusieurs personnes peuvent travailler sur le même compte

(*login*) Unix en même temps. Ces particularités expliquent le fait qu'UNIX soit le système majoritairement utilisé au CERN.

- Ensuite, Unix permet de gérer l'accès aux ressources et aux unités de stockage. En effet, le CERN limite, suivant le type de personnel, l'accès aux ressources et aux données.
- Au même titre que les systèmes Windows, il permet la communication entre les utilisateurs via les moyens modernes classiques. C'est à dire toute les technologies liées à internet : du courrier électronique au protocole FTP (File transfert protocole).
- Enfin, il est aussi environnement de programmation. Les informaticiens du CERN, d'ailleurs, fonctionnent majoritairement sous plate-forme Unix car ce système est beaucoup plus tolérant que les produits que Microsoft en ce qui concerne les langages de programmation.

3.2.1.3 Les quelques commandes Unix au cours de ce stage.

Après avoir entré nom et mot de passe à la suite du message de *login*, à la fin de la procédure d'initialisation de la session de travail, il fallait obligatoirement connaître quelques commandes Unix pour pouvoir travailler. En effet, à la différence de l'environnement Windows dans lequel tout est visuel et cliquable, Unix nécessite encore que l'utilisateur lui lance des ordres textuels.

Il faut relever que ces commandes sont cumulables. Plus elles sont longues, plus la tâche que l'ordinateur doit exécuter est précise.

Par exemple, la commande **ls** crée la liste de ce que possède un dossier, la commande **ls -l** permet de voir ce qu'il y a dans un dossier avec les caractéristiques principales de fichiers (date, taille, ...) et la commande **ls -lrt** permet d'obtenir une liste des fichiers contenus dans un dossier avec leurs caractéristiques mais opère un tri par date.

Ceci peut s'avérer assez fastidieux pour un utilisateur récent. L'impression d'un fichier nommé « bal » sur une imprimante nommée «311» se traduira par exemple ainsi : **XPrint -P 311 bal** . Le problème des commandes Unix est qu'une simple erreur d'espace ou de majuscule empêche l'exécution de l'ordre.

Voici ci-après les commandes essentielles pour pouvoir être fonctionnel sous Unix.

ls -l : liste le contenu d'un fichier avec le détail de son contenu

cd : au début du nom d'un fichier l'ouvre. Seule, elle ferme tous les fichiers

rm : au début du nom d'un fichier, il l'efface

cp fichier dossier : fait la copie d'un fichier dans un dossier

mv fichier1 fichier2 : renomme le fichier1 en fichier2

mv fichier dossier : déplace un fichier dans un dossier

mkdir dossier : crée un répertoire nommé dossier

rmdir dossier : efface un répertoire nommé dossier

man : avant le nom d'une commande, propose une aide pour l'utilisation d'une commande. Seule, elle permet d'obtenir un manuel des commandes Unix.

more man : permet d'obtenir le manuel des commandes en entier.

3.2.2 Emacs

Emacs est un éditeur de texte. Un éditeur de textes sert à créer ou à modifier des fichiers textes qui peuvent être des documents ou des langages de programmation (cf. l'exemple avec les éditeurs html)

Emacs est l'éditeur de texte phare des plate-formes Unix. Je l'ai largement utilisé au cours de ce stage. A première vue, son aspect général est un peu rebutant à coté de l'interface visuelle agréable d'un Word de Microsoft ou même de tout autre éditeur. Cependant, après consultation de ses possibilités dans quelques manuels de vulgarisation qui lui sont entièrement consacrés, force m'est de constater que c'est un outil extrêmement puissant. Ce produit d'UNIX, entièrement libre de tout droit, peut être modifié et étendu à volonté par qui le veut et le peut.

Hormis l'édition de texte, il sert aussi à programmer, aux lectures de news et de courrier électronique, il permet la gestion de fichiers, à envoyer des fichiers par FTP, ...

J'ai utilisé uniquement sa fonction édition de texte. Emacs possède une barre de menu, mais, comme la majorité des programmes du monde UNIX, il adore les touches de raccourci⁸ qui permettent de faire gagner du temps à ceux qui les

⁸ Une touche de raccourci permet d'exécuter une fonction d'un programme sans passer par une barre des menus.

connaissent et d'en faire perdre à ceux qui les ignorent ou qui les utilisent par mégarde.

3.2.3 Le LATEX (prononcez LATEK)

Latex est un programme et un langage créé en 1982 par Leslie Lamport. (Cf.annexe 5)

A l'origine, Latex est un traitement de texte utilisé majoritairement sous UNIX par le monde universitaire et la recherche afin de produire des documents de très grande qualité. LaTeX est un outil très puissant par rapport aux autres logiciels de traitement de texte car il possède la particularité de disposer d'un langage propre : Le LaTeX qui lui permet de disposer de nouveaux outils et d'être très paramétrable.

L'utilisation que la bibliothèque du CERN, fait de ce langage est quotidienne car il permet aux documentalistes de transcrire sur l'écran des lettres grecques et les codes scientifiques en plein texte. En fait , le programme Latex n'est pas un traitement de texte mais un formateur de texte. C'est un petit programme de compilation du langage LaTeX permettant de transcrire le LaTeX en texte lorsque l'utilisateur d'Aleph consulte une notice bibliographique depuis le WEB.

En fait, la plupart du temps, un document Latex est généré par un programme de traitement de texte Unix, Emacs par exemple. Le texte est mélangé avec le code associé et enregistré avant d'être décodé par le compilateur LaTeX. Ensuite, j'ai pu le visualiser par un lecteur spécifique, un programme de visualisation.

Latex ne fonctionne pas comme tous les traitements de texte de type Word car il ne permet pas de visualiser le résultat final au cours de la frappe. Le langage HTML fonctionnait de même avant l'arrivée des éditeurs.

Une formule LaTeX type dans une notice bibliographique du CERN est toujours précédée et fermée \$.

Exemple de notice possédant une formule LaTeX :

Exemples simple : π ou encore γ pour obtenir les lettres grecques correspondantes

Autre exemple, pour créer un plus en exposant : le code est un $^{+}$

Pour un indice : le code est $_{+}$

Il ne faut jamais mettre un espace entre les commandes sinon le code LaTeX n'est pas compilé par LATEX.

La bibliothèque du CERN possède une liste d'autorité pour le formatage du Latex. (cf. annexe 5)

3.3 Agiv : un exemple de traitement semi-automatisé des auteurs

Agiv est un programme qui extrait les auteurs d'un fichier postscript. Le format postscript est un format généralement destiné à l'impression. à l'origine, ce programme s'appelait GIVA et a été créé par le Laboratoire DESY à Zeuthen en Allemagne. Il a été par la suite offert à la bibliothèque du CERN afin de perfectionner ses cross-references (fonctionne comme un renvoi « voir aussi » dans un thésaurus) avec les noms d'auteurs Russes qui possèdent de nombreuses translittérations suivant les pays. Le programme Agiv fonctionne sous UNIX et utilise le protocole de communication TELNET. Agiv cherche de façon automatique le numéro de rapport du document qu'on lui fournit à deux endroits : dans la base de donnée générale du CERN et dans un dossier particulier nommé GIV.

Par défaut, le programme choisi d'ouvrir le fichier postscript situé dans le dossier GIV, car ce dernier permet au documentaliste d'exécuter localement le traitement des auteurs pour un fichier postscript absent de la base.

Une fois le fichier postscript localisé, le programme Agiv l'exporte dans un fichier temporaire nommé AUI où il est retranscrit au format texte et mis en forme avec le champ AU des auteurs en début de chaque ligne.

Après ce traitement, les documentalistes ouvrent le fichier AUI avec emacs et extraient les auteurs du texte principal. En utilisant Aleph en parallèle sur UNIX (système multitâche), les documentalistes peuvent ensuite modifier le champs des auteurs.

3.4 Elaboration d'un petit dossier sur les agents intelligents pour la présentation de copernic99

Nous commencerons cette partie par la définition que donne la lettre de l'Urfist des agents intelligents : « on parle d'agents intelligents lorsqu'on désigne des systèmes qui opèrent dans un environnement qui évolue de manière constante et à propos duquel ces systèmes possèdent une information partielle ou incorrecte. »

L'association française de normalisation (AFNOR) définit ainsi les agents intelligents : « Objet utilisant les techniques de l'intelligence artificielle : l'agent intelligent adapte son comportement à son environnement et en mémorisant ses expériences, se comporte comme un sous-système capable d'apprentissage : il enrichit le système qui l'utilise en ajoutant, au cours du temps, des fonctions automatiques de traitement, de contrôle, de mémorisation ou de transfert d'information.

Un agent intelligent contient un ou plusieurs des éléments suivants :

- Une base de connaissance prédéfinie,
- Un moteur d'inférence, lui permettant de tenir des raisonnements plus ou moins complexes,
- Un système d'acquisition de connaissances,
- Un mécanisme d'apprentissage. »

Le but des agents intelligents est d'optimiser la recherche sur internet. Comme il fallait présenter au groupe l'outil copernic99, issu de ces technologies il m'a semblé utile de fournir un petit dossier annexe sur les agents intelligents et les concurrents de copernic99.

Ce petit dossier s'avérait être une synthèse de l'excellent site de Cybion.fr consacré à la recherche et au développement des techniques concernant les agents intelligents, et d'un très bon dossier élaboré par Béatrice Piau de l'université d'Angers.

Conclusion

Ce stage de fin de formation sous la direction de Madame Ingrid Geretschläger fut pour moi une expérience des plus enrichissante dans le cadre d'une équipe compétente, sympathique et dynamisante. Durant ces 5 mois, j'ai pu voir ce que le travail de documentaliste implique concrètement et me familiariser avec les techniques documentaires les plus avancées.

J'ai eu la chance de pouvoir élaborer en partie un projet ambitieux ; le développement de l'importation automatique des notices bibliographiques.

J'ai eu aussi l'opportunité de pouvoir collaborer avec différents étudiants d'autres universités et d'écoles. Pouvoir échanger avec eux différentes techniques permettant d'amener un projet à son terme fut pour moi extrêmement motivant. Géré sur le long terme, ce projet oriente la bibliothèque du CERN résolument vers une option « tout numérique ». Dans le cadre de ce projet, j'ai pu constater que pour atteindre un objectif, il ne faut pas hésiter à explorer tous les chemins, à faire preuve de souplesse. J'entends qu'il faut savoir accepter les changements et pouvoir réorienter ses directions de recherche. Surtout lorsqu'on est amené à utiliser les technologies nouvelles alors que l'informatique n'est qu'une branche de notre spécialité.

J'ai noté que la bibliothèque, en tant qu'entité administrative se trouve dans un contexte hiérarchique qui peut s'avérer être un frein ou un moteur pour un projet.

Pour finir, je pense qu'à moyen terme, la bibliothèque du CERN devrait installer un système de veille sur les agents intelligents, car ces outils proposent de nombreuses solutions à ses problèmes spécifiques.

Ceci pourrait d'ailleurs être un très bon sujet de stage.

INDEX

A

ACR2	14
Agiv.....	68
Aleph.....	7
Alerte.....	46
ALICE.....	8
annuaire.....	36
APPLETS.....	45
ATLAS.....	8

B

B.E.B.C.	6
BCC.....	50
Berners-Lee.....	15
Bosons B	8
<i>by-passed</i>	13

C

CGI-BIN	42
CMS	8
copernic99.....	3, 68, 69
Copernic99.....	38
<i>Cross-reference</i>	63

D

Delphi.....	7
des langages de haut niveau	44
<i>DESK</i>	10
DSI	11

E

éditeur	51
Emacs.....	14, 66
Ex Libris.....	13

F

FTP.....	52
----------	----

G

<i>gray-book</i>	28
<i>Guestbook</i>	41

I

INSPEC.....	40
ISBD	14

J

JAVA	44
------------	----

L

L3	7
langage de bas niveau.....	44
langage de très haut niveau.....	45
LATEX.....	3, 67
LEP.....	7
LHC.....	7
LHCb.....	8

M

métamoteurs	37
Misk.....	12
moteurs de recherche.....	36

O

<i>offline browser</i>	41
Opal	7
Oracle	14

P

<u>P.S</u>	6
PDF	15
PERL	44
Postscript.....	15
prétirage.....	11

R

<u>recherche directe</u>	62
<u>recherche par index</u>	62
<u>recherche par mot</u>	62

S

<u>S.P.S</u>	6
SÈVE, Roger,	54, 33
SGBD	13

T

Telnet.....	61
-------------	----

U

UNCOVER.....	40
Unix.....	44

V

Visual Basic.....	44
-------------------	----

Bibliographie

Livres :

- Gundavaram, S, La programmation CGI, Sébastopol, O'reilly, 1996
- Horstmann, C.S., Cornell, G ,Au cœur de JAVA, London, The sun microsystems press, 700 p.
- Sallatin, J, Les agents intelligents,Paris, Hermes science publication , 1997, 304 p.
- Moteurs d'indexation et de recherche, Leloup, C, Eyrolles,1997, 164 p.
- Gosling, J., The JAVA™ Application : Programming Interface, Vol. 2., London, Addison-Wesley, 1996.
- Schwartz R. L., Learning PERL., Sébastopol, O'Reilly, 1997.

Périodique :

- Annual report 1998*, Genève, CERN, 1999.
- ARCHIMAG n° 117, septembre 1998, p.41-46
- " *Experiments at CERN in 1998* ", organisation européenne pour la recherche nucléaire, Geneva, 1999.
- Goossens M., " Afficher des documents scientifiques sur le WEB ", Cahiers GUTenberg, n°28, mars 1998, p.181-196.

Electronique :

- Gray-book du CERN sous sa version électronique, <http://www.cern.ch/>
- Mail de Natalie Bouchet et Ingrid Geretschläger sur le site de l'ADBS, <http://www.ADBS.com>

Encyclopédies :

- Encyclopaedia Universalis, 1995
- Grand Dictionnaire Encyclopedique Larousse, 1982.
-

Annexes :

- 1- Listing du personnel du service des prêtirages et des archives de la bibliothèque
- 2- Les sources de la section de gestion de documents
- 3- Mail groupé aux webmestres de l'expérience Opal
- 4- Renvois pour les titres dans la base PREP des périodiques
- 5- Liste d'autorité pour le formatage du LATEX
- 6- Correspondance avec monsieur Schlagheck, webmestre de l'expérience C
- 7- Récupération selon une notice minimale de publication diffusée sur le WEB
- 8- Etat d'avancement de l'aide à la publication sur internet et de soumission simplifiée lors de son arrêt.
- 9- Tableau excel 97 contenant les informations recueillies après une recherche concernant les publications d'Atlas.
- 10- Scripts en langage C et commentaires du programme d'Alerte.
- 11- Correspondance concernant le projet d'Alerte
- 12- Phase finale de l'Alerte.
- 13- Liste des différentes bases de données d'Aleph