Automatisation partielle du traitement de la littérature grise dans le service d'information scientifique du CERN

Catherine Deroche

11 septembre 1998

Auteur : Catherine DEROCHE

Titre : Automatisation partielle du traitement de la littérature grise dans le service d'information scientifique du CERN

Résumé : L'automatisation partielle du traitement de la littérature grise est un projet inéluctable dans un service d'information scientifique aussi important que celui du CERN. Ce rapport présente l'automatisation du rapport annuel par la détection et l'importation automatique d'informations de publication pour les articles écrits par les physiciens travaillant au CERN et par d'autres utilisant les outils du CERN. La recherche de ces articles est effectuée dans la base de données scientifiques INSPEC.

Mots-clefs : Publication préalable - Littérature grise - Conception programme - Traitement automatisé - Interrogation base donnée

Title : Automated processing of grey literature in the CERN's Scientific Information Service.

Abstract : Automated processing of grey literature is essential for a scientific information service as important as the CERN. This report presents the Annual Report automated processing with the detection and importation of publications. This information comes from articles written by physicists who work at CERN and by others who use CERN's facilities. Research for these articles is done by using a scientific database called INSPEC.

Keywords: Preprint - Grey literature - Program design - Automated processing - Database query

Remerciements

Je souhaite remercier toutes les personnes qui de près ou de loin m'ont aidée lors de mon stage.

Je suis très reconnaissante envers Monsieur Corrado Pettenati de m'avoir permis d'effectuer mon stage de fin d'études dans son service, m'offrant ainsi une expérience professionnelle très enrichissante.

Je tiens à assurer ma profonde gratitude à mon maître de stage, Mme Ingrid Geretschlager, pour son aide, son soutien et sa gentillesse tout au long du stage.

Je remercie également Jean-Yves Le Meur pour son aide précieuse, ses conseils, son *support* et pour tout...

Un grand merci à Catherine Cart et Jocelyne Jerdelet pour leur accueil chaleureux dans leur bureau, pour leur aide et leur gentillesse...

Merci à Tullio pour ses conseils, sa disponibilité et sa bonne humeur.

Je remercie aussi tous les membres du service pour avoir bien voulu répondre à mes questions et pour leur aide. En particulier :

- Catherine Bulliard
- Eliane Charney
- Caroline Christiansen
- David Dallmann
- Ruth Eisenberg
- Lamia Hedayati
- Anita Olofson
- Mario Pellacani
- Maiko Real
- Jens Vigen

Merci à Isabelle et Rose pour leur complicité et pour tout...

Note concernant la rédaction du mémoire

Pour simplifier la compréhension, tous les termes anglais ont été typographiés en *italique* et les commandes de programmation en style `machine à écrire`.

Les illustrations et les différents programmes ont été placés en annexe.

Pour distinguer les notes de bas de page des renvois bibliographiques, ces derniers ont été placés entre crochets.

Table des matières

Introduction

Ce mémoire illustre le stage de fin d'études effectué dans le cadre du DESS Informatique Documentaire de l'ENSSIB (Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques) et de l'université Claude Bernard de Lyon 1 du 1 Juin au 30 Septembre 98. Il présente le travail que j'ai réalisé au Laboratoire Européen pour la Physique des Particules, CERN, à la division du Support Administratif, AS, sous la direction du Mme Ingrid Geretschlager, Dr. et conservateur ENSSIB, responsable de la section du traitement documentaire.

Le CERN, un des plus grands centres de recherche en physique des particules, publie chaque année un rapport constitué de trois volumes dont l'un est élaboré par le service d'information scientifique et concerne la liste des publications du CERN. Ce rapport est effectué par l'équipe des pré-tirages et localise les références bibliographiques complètes des résultats des activités du CERN.
Il reflète la production intellectuelle des équipes de chercheurs du CERN et joue un rôle essentiel dans l'établissement d'une politique budgétaire.

Ma mission a été de proposer une solution pour automatiser une partie de ce rapport, en détectant et en important partiellement dans la base de donnée du CERN des références d'articles de la base de données scientifiques INSPEC.

Ce mémoire, constitué de 3 parties, présente dans une première partie le contexte du projet. La seconde partie développe la réflexion et la méthodologie adoptée sur le travail effectué. L'ensemble des autres activités est étudié dans la dernière partie.

Chapitre 1

Le CERN

1.1 Présentation

Le CERN dont l'acronyme signifiait à l'origine Conseil Européen pour la Recherche Nucléaire est devenu le Laboratoire Européen pour la Physique des Particules. Fondé en 1954, il est une des premières collaborations européennes et représente un exemple brillant de coopération internationale [1].

A l'origine, 12 états membres ont signé sa convention constitutive et actuellement, le CERN compte 19 pays [1] membres.

Le Laboratoire se situe à la frontière Franco-Suisse à l'ouest de Genève au pied des montagnes du Jura.

L'activité du CERN est la physique théorique, celle qui explore les structures de la matière et veut répondre aux questions existentielles que chacun a pu se poser telles que :

- Qu'est-ce que la matière ?
- D'où vient- elle ?
- Comment la nature opère-t-elle pour maintenir une telle cohésion entre ces atomes ?

Pour répondre à ces questions, cinq accélérateurs de particules ont été construits : ils permettent de sonder le coeur de la matière grâce à des faisceaux

[1] Autriche, Belgique, République Tchèque, Danemark, Finlande, France, Allemagne, Grèce, Hongrie, Italie, Pays-Bas, Norvège, Pologne, Portugal, République Slovaque, Espagne, Suède, Suisse et Royaume Uni

de particules de hautes énergies. Ces accélérateurs sont de plus en plus performants et permettent d'atteindre des énergies de plus en plus élevées : de 600 MeV pour le synchro-cyclotron jusqu'à 14 TeV pour le futur LHC.
Ce dernier sera fonctionnel en 2002 et permettra de recréer les conditions régissant l'univers 10 secondes après le Big Bang et ainsi d'en savoir un peu plus sur notre passé.

Les découvertes faites au CERN ne sont pas purement théoriques mais permettent de développer et de concrétiser de nouveaux produits et nouvelles technologies tels que le WEB, l'imagerie médicale, des puces électroniques perfectionnées....

Environ 6 500 physiciens, soit la moitié des scientifiques de la physique des particules, utilisent les installations du CERN ce qui représentent 500 universités et plus de 80 nationalités.

En 1997, le budget global du CERN était de 877 millions de francs suisse.

Devant l'ampleur de la mission du CERN, le service d'information scientifique se doit d'être le plus performant et le plus perfectionné possible. C'est d'ailleurs au CERN qu'est né le WEB, symbole des nouvelles technologies de l'information : il répondait à l'origine aux besoins des chercheurs de diffuser la littérature grise à travers le monde.

1.2 Le service d'Information Scientifique

Le service d'Information Scientifique réunit une bibliothèque centrale, cinq bibliothèques satellites et des archives historiques et scientifiques. Il a pour mission d'acquérir et de gérer l'information concernant les travaux du CERN et de les diffuser à la communauté scientifique.
Le service est divisé en plusieurs sections :

- Gestion des documents

  Ce service a pour fonction de cataloguer et de tenir à jour l'ensemble de la base de données de la bibliothèque et plus particulièrement les pré-tirages. Les pré-tirages ou *preprint* sont les articles que les physiciens soumettent à un journal avant publication et sont diffusés comme littérature grise. Il peut s'écouler une période de 6 mois à 2 ans entre la pré-publication et la publication.
  Il élabore aussi le volume III du rapport annuel du CERN [2,3] et

produit une liste hebdomadaire de pré-tirages. Il gère également l' acquisition de monographies.

- Gestion des périodiques

  Ce service gère les revues papier et électroniques. L'accent est mis actuellement sur les périodiques électroniques qu'il faut faire découvrir aux lecteurs. Pour cela, toute une 'campagne publicitaire' est organisée.

- Service aux utilisateurs

  Il s'occupe des prêts inter-bibliothèques, de la gestion des pages WEB de la bibliothèque, de la distribution des publications CERN...

- Les archives

  Le catalogage, la recherche de documents et la remise en état d'ouvrages sont les principales activités de ce service qui doit garantir la sécurité des documents.

En 1997, le service d'information scientifique a bénéficié d'environ 1 million de francs suisses dont la moitié a été attribuée à l'acquisition de périodiques. L'achat des monographies correspond à un cinquième du budget des abonnements.

Quelques chiffres

- 200 000 pré-tirages présents dans le catalogue du CERN
- 20 000 pré-tirages acquis en 1997
- 80% de l'information est de la littérature grise en anglais
- 40 000 monographies
- 1 000 titres de périodiques
- 180 abonnements à des journaux électroniques
- 7 500 prêts par an

1.3 Aleph ou la gestion électronique des documents

1.3.1 Présentation d' Aleph

Aleph (*Automated Library Expandable Program*) est un logiciel de gestion des bibliothèques et des centres de données, développé par l'université de Jérusalem et produit par la société Ex-libris [4,5].

C'est un système de gestion des documents qui peut être adapté à différents types d'instituts : bibliothèque, musée, archive, centre de recherche.
Aleph permet de traiter différents types de documents : livre, article, rapport, carte, publication, schéma, brevet, microfiche et ce, avec une grande quantité de caractères comme les alphabets latin, arabe, hébreux et mathématique. Il est donc très adapté à la structure du CERN.

Il permet une gestion très flexible et peut être modifié en fonction des besoins de chaque utilisateur : le nombre de champ pour un enregistrement n'est pas limité ainsi que la longueur du champ, le format d'entrée est fixé par l'utilisateur.
Aleph développe de nombreuses fonctions : acquisition, recherche, catalogage, circulation, gestion des prêts entre bibliothèques, budget, édition de listes...
Aleph utilise le *Common Command Language* sous Unix.

1.3.2 Structure d'Aleph

Aleph, du fait de sa structure très souple, a été adapté aux différents types d'information présents à la bibliothèque.

Chaque type de document a une structure spécifique [2](les champs diffèrent selon les besoins) et possède une base propre. La base de données globale du CERN est, en fait, constituée de plusieurs sous-bases qui sont indépendantes. Par exemple, la base 11 est réservée aux pré-tirages, la base 12 aux conférences, la base 13 aux articles publiés.
Les recherches peuvent donc se faire soit sur la base globale soit sur une base particulière.

L'accès à Aleph peut se faire par Telnet (réservé aux catalogueurs) ou par

[2]cf. exemples de documents de la base ALICE annexe A

une interface WEB, accessible au grand public `http: //alice.cern.ch/`.

Chapitre 2

Mise en place d'une procédure automatique pour le traitement du rapport annuel

2.1 Objectifs

Le projet concerne une publication de l'équipe des pré-tirages, le rapport annuel, volume III. Ce rapport répertorie les références bibliographiques des articles publiés concernant le CERN. Ces articles sont de deux sortes : des *CERN Works* ou des *CERN Papers*.

Les *CERN Papers* sont des publications écrites par des chercheurs du CERN alors que les *CERN Works* sont écrits par des chercheurs qui utilisent les ' machines' du CERN mais qui ne sont pas rattachés au CERN.
Le service d'Information Scientifique du CERN ne reçoit que 50 % de ces pré-tirages qui sont soit envoyés à la bibliothèque, soit obtenus sur le serveur de Los Alamos, autre laboratoire de la physique des particules.
La liste des publications du CERN 1997 contient plus de 1800 références bibliographiques.

Les chercheurs sont invités à soumettre leurs articles à la bibliothèque [1] mais beaucoup 'oublient'. Tous les pré-tirages qui n'ont pas été soumis au CERN doivent être recherchés dans les bases de données extérieures : ce sont les *By-Passed*.
Jusqu'à présent, ce travail est entièrement manuel et représente une charge

[1] d'après la circulaire numéro 29, document officiel stipulant les règles CERN en matière d'information scientifique auxquelles les auteurs doivent se soumettre,cf annexe B

importante pour le service. Il semble indispensable de l'automatiser vu l'augmenta tion constante du nombre d'articles (+ 60 % entre 1990 et 1997).

Mon projet de stage est donc de proposer une procédure pour automatiser le traitement du rapport annuel afin d'alléger le travail du service et à moyen terme de permettre le développement du fonds documentaire.
Il s'agit d'un travail de réflexion en équipe afin d'élaborer différents scénarios et voir quelle est la meilleure solution. Mais, c'est aussi un travail de conception puisque j'ai dû réaliser le programme permettant de détecter ces *By-Passed* et de les importer ou bien d'importer uniquement l'information de publication dans le cas d'un non *By-Passed*. Pour mon travail, il est essentiel de distinguer les articles qui sont déjà dans la base du CERN et les *By-Passed*.

2.2 Le Rapport Annuel avant

Les sources d'information du service

Des statistiques [2] sur le rapport annuel, vol. III viennent d'être réalisées sur les données entrées dans la base du CERN.
Environ 25 % des chercheurs du CERN font part à la bibliothèque de leurs prétirages. L'information externe provient essentiellement de Los Alamos (pour 25 %) qui est un autre laboratoire de la physique des particules. Les 50 % restant sont des articles trouvés 'manuellement' donc des *By-Passed*.

Le traitement de l'information

Jusqu'à présent, environ 10 personnes ont travaillé sur ce rapport ce qui correspond à deux personnes à temps plein soit un coût de 10 000 ChF (estimation moyenne).

Les références de publication sont recherchées manuellement lors du dépouille ment des comptes rendus de conférences et des périodiques qui sont acquis par le CERN. La détection des articles publiés du CERN est donc très aléatoire. Le dépouillement permettait de localiser les articles pour lesquels la bibliothèque n'avait jamais obtenu de pré-tirage. INSPEC servait à vérifier occasionnellement certaines informations.
La notice bibliographique des *By-Passed* est ensuite saisie manuellement.

[2] cf. annexe C

L'année dernière, environ 650 *By-Passed* ont été trouvés ainsi.

2.3 Principe de l'automatisation

2.3.1 Analyse des besoins

Plusieurs étapes sont à envisager. Il s'agit dans un premier temps de détecter tous les *CERN Papers* et *CERN Works* et ainsi de diminuer le caractère aléatoire de leur découverte. Puis, dans un deuxième temps,

* si ces notices sont des *By-Passed*, il faut les importer dans la base de données du CERN, qui s'appelle *ALICE*. Pour cela, un reformatage des données est nécessaire.

* si ce ne sont pas des *By-Passed*, les notices sont donc dans ALICE mais sont peut-être encore à l'état de pré-tirages c'est à dire non publiées. Dans ce cas, il faut importer l'information de publication et leur attribuer une autre sous-base (celle des articles publiés).

  De plus,

  - si l'article provient d'un périodique, il faut modifier le champ PR (*Publication Reference*)
  - s'il provient d'une conférence, il faut ajouter un champ LKR (*Link Retrieve Reference*)

2.3.2 Utilité de INSPEC dans la procédure d'automatisation

Présentation de INSPEC

Cette banque de données bibliographiques (`http: //www.iee.org.uk/publish/inspec/`), produite par l'Institution of Electrical Engineers (IEE) depuis 1969, couvre de nombreux domaines scientifiques : la physique, l'électronique, l'ingénierie nucléaire, l'ingénierie électrique, l'informatique, l'intelligence artificielle, les technologies de l'information.

INSPEC dépouille notament :

- Computer and Control Abstracts
- Physics Abstracts

- Electrical and Electronics Abstracts

La langue d'interrogation est l'anglais.

INSPEC scanne 4 100 périodiques dont 750 sont résumés [6]. Cela correspond à 82% de la base INSPEC. Le reste provient de compte- rendus de conférences (16%), livres (1 000 par an environ), rapports, thèses, brevets...

Depuis 1969, plus de 5 850 000 notices bibliographiques ont été incorporées dans la base. La mise à jour est hebdomadaire. La base INSPEC s'accroît de 11 000 références tous les 15 jours.
INSPEC fête cette année ses 100 ans [7] ; en réalité, c'est le centenaire de la publication du premier numéro de *Science Abstract*, journal à partir duquel a été développé la base de donnée.

Couverture de la base du CERN par INSPEC

Pour savoir si la base de données INSPEC était la plus adaptée pour ce projet, il a fallu étudier la couverture de la base du CERN par INSPEC en comparant la liste des périodiques CERN avec celle d' INSPEC. Ce travail a été effectué par Alexandre Hundzinger, un jeune stagiaire.

Il s'est avéré que 66 % des périodiques CERN sont dépouillés par INSPEC ce qui est tout à fait satisfaisant.
Par la suite, deux listes ont été établies : une des périodiques absents de INSPEC qui seront traités manuellement et une autre contenant les titres couverts par INSPEC qui seront traités automatiquement.

2.3.3 Réflexion pour la mise en place d'une procédure automatique

Le premier problème à résoudre a été l'obtention de l'ensemble des références des *CERN Papers* et des *CERN Works*.

INSPEC couvrant bien les domaines scientifiques étudiés par le CERN, nous avons donc décidé de l'utiliser comme source d'information.
Le service d'information scientifique a accès à INSPEC par le WEB ou par Telnet par le biais d'un abonnement à DIALOG.

Il était difficile d'automatiser la procédure en utilisant l'interface WEB car celle-ci n'offre pas la possibilité de sauvegarder les équations de recherche. L'accès par DIALOG bien que plus cher offre la possibilité de sauvegarder les équations de recherche, de combiner les requêtes facilement et de choisir le format des notices bibliographiques obtenues. Nous avons donc décidé d'interroger INSPEC par DIALOG.

Le problème suivant a été de définir la stratégie de recherche. Deux possiblités existaient : soit on se calquait sur la procédure manuelle et on interrogeait INSPEC pour chaque compte rendu de conférence à dépouiller, soit on élaborait une équation enveloppant toutes les comptes rendus.
Nous avons adopté la deuxième solution plus performante. Elle permet de rechercher les articles CERN avec seulement deux requêtes. L'organigramme de cette procédure est présenté ci-après (Fig. 1).

Catherine Cart, documentaliste, avait établi une liste de mots-clés extraites de l'index du dernier rapport annuel, vol. III, utilisée pour l'interrogation manuelle de INSPEC mais après quelques essais, il est apparu qu'elle apportait du bruit. Je l'ai donc modifiée et elle pourra être complétée à tout moment, par exemple, si une nouvelle expérience est créée au CERN.

La procédure automatique place donc INSPEC en début de chaine alors que dans la procédure manuelle testée peu de temps auparavant, INSPEC se situait en fin de chaine.

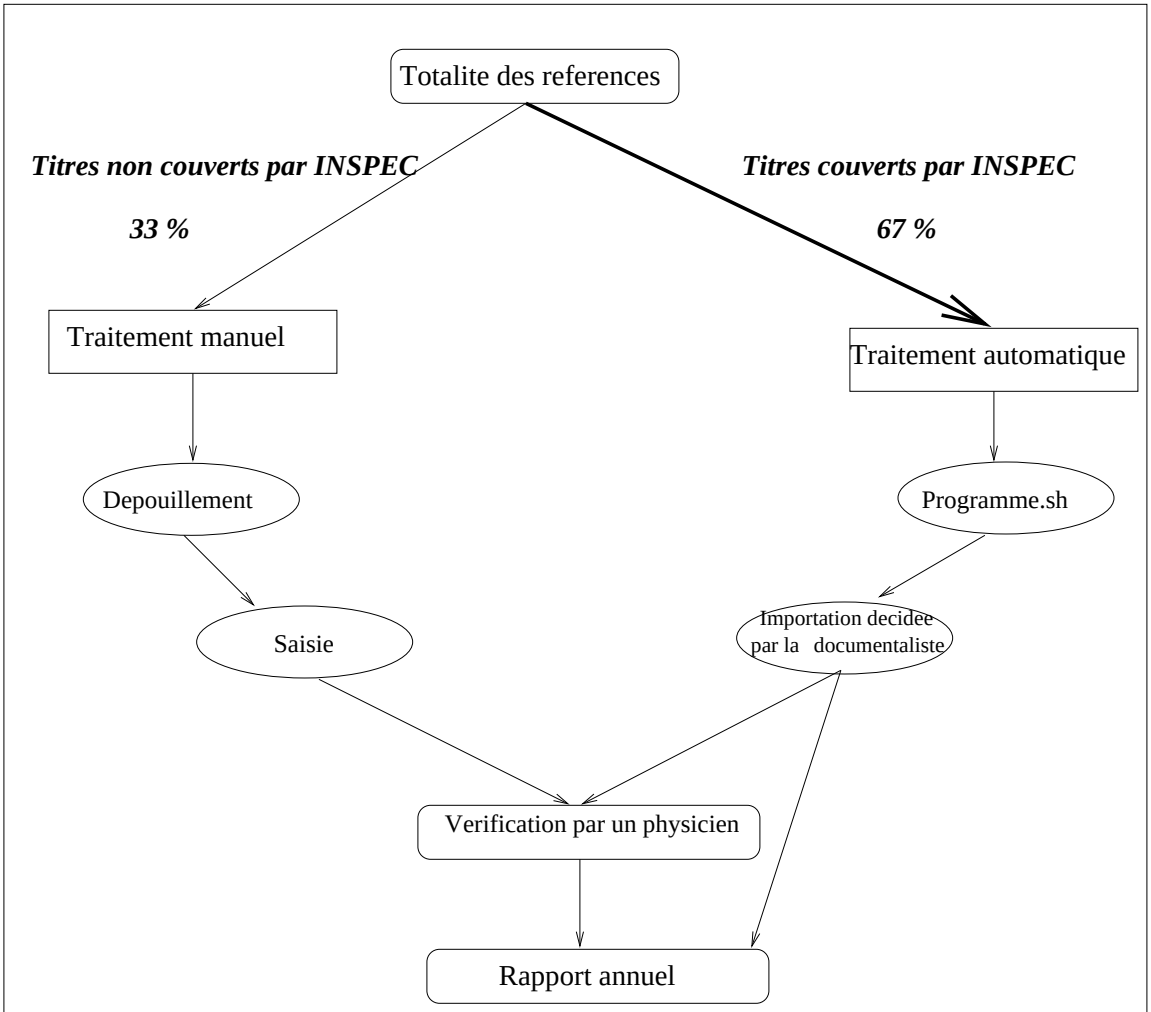

Fig. 1 : Organigramme du traitement general du rapport annuel

La deuxième étape a été de définir les fonctions du programme. J'ai alors élaboré l'organigramme (Fig. 2) qui résume la démarche.

Nous avons également réfléchi sur les besoins exacts pour savoir quelle information devait être importée ou non.

Finalement, les champs à extraire lors de l'importation totale de la notice (*By-Passed*) sont :

- le titre
- les auteurs
- le numéro INSPEC
- l'affiliation
- les informations concernant la conférence
- le titre du périodique
- le volume
- la pagination
- l'année de publication
- Il faudra aussi ajouter des informations intrinsèques à ALICE comme le numéro de la base, le champ `SW` *Status Week* ou le champ `NI` *Note Interne*.

Pour l'importation de l'information de publication, seuls le titre du périodique, le volume, la pagination et l'année de publication doivent être importés.

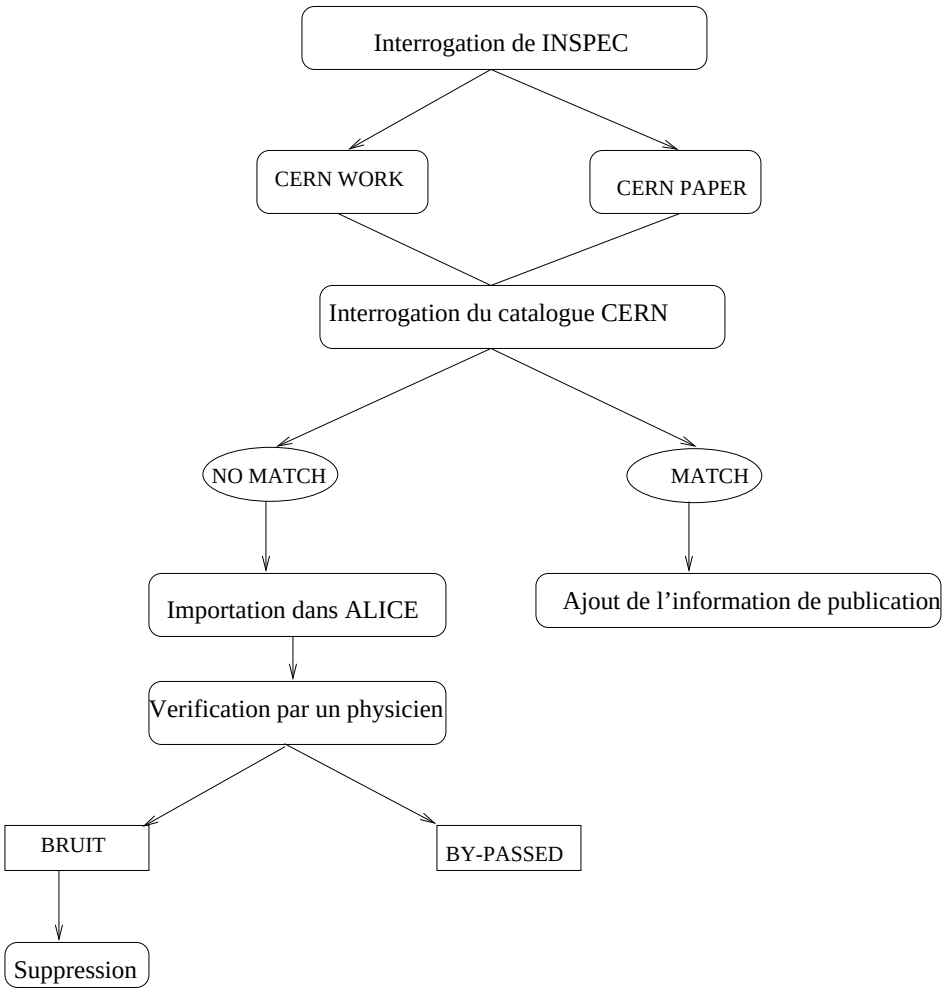

Fig. 2 : Organigramme de la procedure d'automatisation du rapport annuel

2.4 Mise en place d'une procédure automatique

2.4.1 Présentation du projet

Ce rapport[3] préliminaire *Automatisation du traitement annuel* a été rédigé dans le but de présenter à M. Pettenati, directeur du Service Scientifique d'Information les enjeux de cette automatisation.

La rédaction de ce document m'a permis de faire la synthèse du travail à effectuer mais aussi d'apprécier les aspects financiers et humains. Cette partie a été difficile à évaluer mais Mme Geretschlager m'a beaucoup aidée en me donnant les estimations du temps de travail sur le rapport annuel de chaque personne de l'équipe ainsi que différents coûts.

Ensemble, nous avons évalué ce qu'entraînerait une automatisation du rapport annuel comme par exemple le gain de temps, estimé à un poste temps complet.

2.4.2 Elaboration d'une équation de recherche

L'équation de recherche a été élaborée de manière à retrouver les *CERN Papers* et les *CERN works*.

La première étape est la recherche des publications entrées dans INSPEC durant le mois écoulé. Tullio Basaglia, bibliothécaire détaché de l'institut polytechnique de Turin, qui connait très bien DIALOG et gère les abonnements aux différentes bases de données, m'a conseillé sur la méthode à suivre et pour tout ce qui concerne les différents tarifs de DIALOG.

Comme nous avons donc choisi de ne pas utiliser le titre de la conférence pour trouver les *By-Passed*, nous sélectionnons les articles qui ont été entrés dans la base INSPEC durant le mois précédent pour faire une pré-sélection des articles récents. DIALOG possède une commande `UD` (*Updated Date*) qui le permet facilement.

Puis, nous recherchons les publications qui ont une affiliation CERN grâce au champ `CS` (*Corporate Source*) qui correspond à l'affiliation. Ainsi, nous

[3] cf. annexe D

détectons les *CERN Papers*.

Le plus difficile a été d'établir des mots clés pertinents qui permettent de retrouver les publications identifiant le CERN. Pour cela, nous avons extrait de l'index [4] du dernier rapport annuel le nom de toutes les collaborations, de tous les accélérateurs et de toutes les expériences. Cela recouvre bien l'activité du CERN et permet d'identifier les *CERN Work*.
L'équation obtenue est conséquente (environ 1500 caractères), mais tous les termes semblent pertinents.

Principe de la recherche

1. **Recherche des dernières publication**

   UD = AAMM Ex : mai 1998 $\rightarrow$ 9805

2. **Recherche des publications avec affiliation CERN**

   CS =cern

3. **Recherche des publications avec la liste des mots clés identifiant le CERN dans les champs titre, résumé, mots-clés**

```
((AD()FEASIBILITY()STUDY()TEAM) OR ALEPH OR ALICE
 OR APE OR ASACUSA OR ATHENA OR ATLAS OR (BEATRICE
()COLLABORATION) OR (CERES()COLLABORATION) OR CERN
 OR (CHARM()COLLABORATION) OR (CHORUS()COLLABORATI
ON) OR CLIC OR CMS) OR  ((COMPASS()COLLABORATION)
 OR COSMOLEP OR (CPLEAR()COLLABORATION) OR DELPHI
 OR (E()632) OR EHS OR (EMU()01) OR FELIX OR (GAMS
()COLLABORATION) OR HEC)  OR ((IHEP()IISN()LANL()
LAPP) OR ISOLDE OR JETSET OR ((LEBC()EHS)) OR LEP
 OR (LINEAR()ELECTRON()PROTON) OR LHC OR (LARGE()
HADRON()COLLIDER) OR LHCB OR L3 OR MACRO OR MIC
 OR (MISTRAL()COLLABORATION) OR (MOOSE()COLLABORA
TION) OR (NA()22) OR (NA()35)) OR  ((NA()36) OR
(NA()38) OR (NA()44) OR (NA()45) OR (NA()47) OR
(NA()48) OR (NA()49) OR (NA()50) OR (NA()51) OR
(NA()52) OR (NA()56) OR (NEW()MUON()COLLABORA
TION) OR NICOLE OR (NMC()COLLABORATION) OR (NOMAD
()COLLABORATION) OR (OBELIX()COLLABORATION) OR
(OMEGA()COLLABORATION) OR OPAL OR (PROTON()
```

[4] extrait de l'index utilisé cf. annexe E

```
SYNCHROTRON) OR ((PS()185) OR (PS()205) OR (PS
()208) OR (PS()210) OR (PS()212) OR (RD()3) OR
(RD()5) OR (RD()8) OR (RD()19) OR (RD()20) OR
(RD()33) OR (RD()34) OR (RD()36) OR (RD()37) OR
 (RD()40) OR (RD()42)) OR  ((RD()48) OR REX OR
(ROSE()COLLABORATION) OR (SICAPO()COLLABORATION)
 OR (SPIN()MUON()COLLABORATION) OR  SPS OR SPY
 OR (TILECAL()COLLABORATION)  OR TOSCA OR TOTEM
 OR WAL OR (WA()56) OR (WA()80) OR (WA()85) OR
(WA()89) OR (WA()94) OR (WA()97) OR (WA()98)
OR (WA()102))
```

4. **Recherche des *CERN Papers***
   1 and 2

5. **Recherche des *CERN Works***
   (1 and 3) not 2

2.4.3 Interrogation d'INSPEC par l'intermédiaire de DIALOG

DIALOG offre la possibilité de sauvegarder les équations de recherche grâce à la commande `save` et de les rappeler par la commande `ex`.

Interrogation de INSPEC

`telnet dialog.com`
login + mot de passe
`?  b4` (INSPEC est la base 4 de DIALOG)
`?  ex sdsdinsp` (appelle l'équation sauvegardée)
`?  s ud=AAMM` (recherche les entrées dans INSPEC du mois)
`?  s s1 and s2` (recherche les *CERN Works*)
`?  t 3/4/all` (affiche les notices de la recherche 3 dans le format 4)
`?  s (s1 and CS=CERN) not s3` (recherche les *CERN Papers*)
`?  t 4/4/all`
`?  logoff`

J'ai choisi d'afficher les résultats dans le format 4 qui donne la notice bibliographique entière avec en plus, des *tags* qui identifient chaque champ. Cela simplifiera la programmation notamment pour l'extraction de l'information des différents champs.

Toute l'interrogation sur DIALOG est sauvegardée par l'intermédiaire de la commande UNIX *script*.

Ce fichier obtenu va servir de point de départ pour le programme.

2.4.4 Analyse du fichier obtenu de INSPEC

Après analyse des premiers documents obtenus, il est apparu que beaucoup de références n'étaient pas très pertinentes.

Les causes de bruit sont notamment dues à l'emploi des nombreuses abréviations et des acronymes.
Par exemple, le terme *PS* choisi car il identifie la division du CERN *Power Supply*, a de nombreuses significations : *PS* est l'abréviation du polystyrène, du format postscript, de *porous silicon*, de la picoseconde...
Omega, aussi, qui est le nom d'une collaboration amène tous les documents ayant une formule mathématique comportant Ω.
Pour élimimer ce bruit, nous avons précisé qu'il s'agissait de collaboration dans le cas où la collaboration n'est pas aussi le nom d'une expérience du CERN.
De 800 notices bibliographiques avec la première équation, nous sommes ainsi passées à 200 notices.

Il est cependant difficile d'évaluer le bruit : en effet, l'activité du CERN est très vaste et seul un physicien connaissant bien les travaux du CERN peut évaluer la pertinence des articles.

2.4.5 Conception d'un programme en shell UNIX

Objectifs du programme

Le point de départ du programme est le fichier obtenu par DIALOG.

Le programme doit donc extraire le titre et le premier auteur de chaque notice bibliographique pour les comparer au catalogue de la bibliothèque du CERN.

Le traitement diffère selon le résultat de cette interrogation.
Si l'article est déjà dans la base sous la forme de pré-tirage, il suffit d'importer les informations de publication.
Par contre, s'il n'est pas dans la base, il faut importer la notice bibliographique entière et créer certains champs.

Toutes ces informations seront vérifiées avant d'être importées dans la base du CERN et regroupées dans une semaine fictive du champ `SW` qui permettra de les traiter.

David Dallman, physicien détaché auprès de la bibliothèque, vérifiera alors la pertinence de ces notices et si les articles conviennent, il les changera de semaine (champ `SW`)et ajoutera des informations concernant la division, l'expérience, la collaboration qui sont des champs propres au CERN. Si les articles n'entrent pas dans le cadre de recherche du CERN, il les effacera.

Démarrage du programme

La programmation a constitué la majeure partie de mon activité.

Le début de la programmation a été un moment assez incertain. En effet, n'étant pas analyste programmeur et ne connaissant pas vraiment de langage de programmation, il a été difficile de démarrer.

Le langage le plus adapté à ce genre de programme est PERL, langage dérivant du C et des scripts Unix. Or, je ne connaissais pas ce langage et il paraissait assez difficile de l'apprendre puis de programmer en trois mois. Après de nombreuses discussions, Jean-Yves Le Meur, ingénieur informaticien chargé du catalogue de la bibliothèque du CERN, m'a beaucoup aidée tout au long de la conception de ce programme et m'a conseillée d'utiliser les scripts Shell. Ce langage est plus simple à apprendre que PERL et est adapté à ce type de développement [8,9,10,11].

Mon environnement de travail était l'environnement standart UNIX du CERN. J'ai travaillé sur un terminal X en me connectant à un serveur SUN.

Quelques commandes Shell indispensables

- *cat :* affiche sur la sortie standard le contenu d'un ou plusieurs fichiers
- *cp :* effectue une copie physique d'un fichier sur un autre
- *echo :* affiche la liste des paramètres sur la sortie standard
- *grep :* sélectionne dans un fichier les lignes contenant une expression particulière
- *man :* accès en ligne au manuel !
- *mkdir :* crée un répertoire
- *mv :* change le nom d'un fichier
- *rm :* supprime un fichier
- *script :* lance un nouveau processus Shell et enregistre dans un fichier la trace de la session

Deux commandes UNIX ont été très utilisées pour le traitement des textes:

- *awk :* utilisé pour sélectionner une ligne qui commence par une expression particulière
- *sed :* transforme une chaine de caractère en une autre, très utile pour le reformatage

Algorithme du programme principal

Le programme principal *main.sh* est en fait constitué de cinq sous-programmes [5] qui correspondent à chaque étape de l'algorithme. Le paragraphe suivant développe en détail chaque programme.

L'algorithme du programme principal[6] est présenté ci-après (Fig. 3).

[5] leurs noms sont écrits en italique

[6] Programme *main.sh*.cf. annexe F

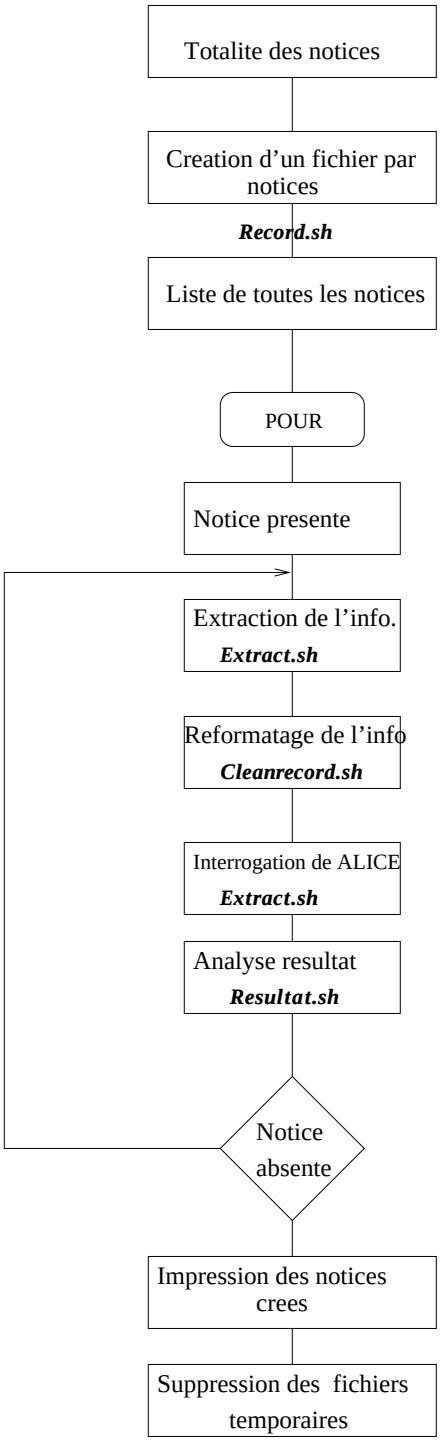

Fig. 3 : Algorithme general du programme main.sh

Détails des programmes

1. RECORD.SH
   Ce programme[7] permet de créer un fichier distinct pour chaque notice bibliographique présente dans le fichier obtenu par DIALOG [8].

   Une notice bibliographique est identifiée par le champ AZ. Chaque fois que le programme rencontre ce champ, il crée un nouveau fichier et place toutes les données qui suivent et ce, jusqu'au prochain champs AZ, dans ce fichier.
   Il crée donc autant de fichiers qu'il y a de notices bibliographiques dans le fichier DIALOG.

   De plus, un premier 'nettoyage' est effectué : les lignes ne contenant pas d'information utile (comme le champ FN ou CZ) sont éliminées. Le champ AZ correspondant à l'identifiant de la notice INSPEC et donc étant unique, il a servi à nommer le fichier contenant la notice.

   Exemple : Contenu du fichier NUMINS5940603 obtenu après cette première opération

```
AZ- 5940603|
TI- Shear tests on adhesives for magnet collars for the LHC|
AU- Tavlet, M.; Hominal, L.|
CS- CERN, Geneva, Switzerland|
JN- Cryogenics|
CP- UK|
VL- vol.38, no.1|
PG- 47-50|
PY- Jan. 1998|
CO- CRYOAX|
SN- 0011-2275|
CD- <US COPYRIGHT CLEARANCE CENTER CODE> 0011-2275/98/$19.00|
CT- Cryogenics (UK)|
CT- International Cryogenic Materials Conference. Topical Conference:
    Nonmetallic Materials and Composites at Low Temperatures VIII|
CL- Geneva, Switzerland|
CY- 23-25 Sept. 1996|
```

[7] cf. annexe G
[8] extrait du fichier DIALOG, cf. annexe H

PU- Elsevier
DT- Conference Paper (PA); Journal Paper (JP)|
LA- English|
TC- Practical (P); Experimental (X)|
MI- C050-98003|
RF- 5|
AB- Thermal tests and radiation tests have been carried out
on four epoxy adhesive systems which cure at room temperature.
Results show that one system is suitable for application at
1.8 K in the radiation environment of the future Large Hadron
Collider to be built at CERN.|
DE- adhesion; shear strength; superconducting magnets|
ID- magnet collars; LHC; adhesives; shear tests; thermal tests;
Large Hadron Collider; CERN; 1.8 K|
NI- temperature 1.8E+00 K|
IC- 0011-2275(199801)38:1L.47:STAM;1-R|
DN- S0011-2275(97)00109-4|
CC- A8190 (Other topics in materials science); A8140L
(Deformation, plasticity and creep)||
CG- Copyright 1998, IEE|

2. EXTRACT.SH

Un programme mkalice.sh déjà existant permet de créer un nouvel enregistrement dans la base de données ALICE. Il faut pour cela que chaque champ de l'enregistrement soit dans un fichier indépendant : par exemple, le titre doit être placé dans un fichier intitulé TI.

Ce programme extract.sh[9] permet d'extraire l'information présente dans chaque champ et de la placer dans un fichier au nom du champ. Il sélectionne aussi le titre et un auteur, les reformate pour permettre d'interroger le catalogue. L'interrogation se fait en extrayant les quatre premiers mots significatifs du titre. Les mots vides tels que a, the, and, for sont élimiés systématiquement.

L'interrogation de ALICE se fait par l'interface WEB du catalogue grâce à LYNX [12], un logiciel offrant la possibilité d'effectuer des opérations sur les adresses http du WEB. En effet, il suffit de définir sur l'interface WEB du catalogue la requête désirée et de la modifier à l'aide de variables.

Interrogation dans Alice par LYNX :
La ligne de commande ci-dessous permet de formuler une requête sur le catalogue du CERN.

```
 lynx -source "http://alice.cern.ch/search/adsearch?uid=der
oche&freetext1=$linetitre&field1=wti&operator=AND&freetext2
=$lineauthor&field2=wau&base=Preprints" > lynx$source
```

L'interrogation se fait en définissant quatre paramètres :

- L'UID (*User Identifyer*)permet de choisir le type de format du résultat de la requête, c'est donc ce qui apparait dans le fichier de sortie lynx$source. Je l'ai défini ainsi :

```
SYSNO: $$SYSNO <BR>
TITRE: $$TI      <BR>
AUTEUR: $$AU    <BR>
PR: $$PR <BR>
LKR: $$LKR <BR>
SW: $$SW    <BR>
```

[9] cf. annexe I

Exemple

```
SYSNO: 0125011
TITRE: Pion-baryon correlations in nucleus-nucleus collisions
       between 400 and 800 MeV per nucleon
AUTEUR: Gosset, J ; et al.
LKR: . Subm. to: 18th International Workshop on Gross
     Properties of Nuclei and Nuclear Excitations Hirschegg,
     Austria ; 15 - 20 Jan 1990 .
     Publ. in: Proceedings H Feldmeier GSI, Darmstadt, 1990
SW: n 9043n
```

- La base sélectionnée est celle des `preprints` qui correpond aux articles publiés et aux articles non publiés.

- Une variable *linetitre* et une variable *lineauthor* permettent d'interroger le champ auteur et le champ titre du catalogue CERN.

3. CLEANRECORD.SH

   Comme son nom l'indique, ce programme[10] 'nettoie' les fichiers avant l'importation. Il teste la présence de certains champs, les transforme ou les crée. Il a été modifié au fur et à mesure des 'bugs' détectés...

   Exemples de modifications sur le champ auteur et sur le champ PR

   - Le champ Auteur dans INSPEC est de la forme :
     AU- Hastings, M.B.; Levitov, L.S.—
     Dans ALICE, il s'écrit :
     AU- Hastings, M B ; Levitov, L S ;
     Il faut donc transformer les '; ' en ' ' et supprimer le symbole '—'.
   - Le champ PR n'existe pas dans INSPEC. Il a donc fallu le créer.
     Dans ALICE, il est de la forme :
     PR- $$pTitre du periodique : Volume (Année)$$cPage

     Dans INSPEC, le titre du periodique est dans le champ JN, le volume dans le champ VL, les pages dans le champ PG et l'année dans le champ PY. Il faut donc construire le champ PR en regroupant toutes ces informations et en les plaçant dans les sous -champs correspondants.

[10] cf. annexe J

4. RESULTAT.SH

   Ce programme[11] analyse le fichier obtenu après l'interrogation dans ALICE. Il permet à l'utilisateur de visualiser les résultats et propose une interface avec différents menus.

   Ce programme en appelle un autre, mkalice.sh, qui crée la notice bibliographique correspondant au formatage utilisé au CERN. Le programme mkalice.sh est utilisé pour d'autres importations dans la base, je l'ai juste adapté en modifiant certains paramètres spécifiques de INSPEC.

[11] cf. annexe K

Algorithme du programme resultat.sh

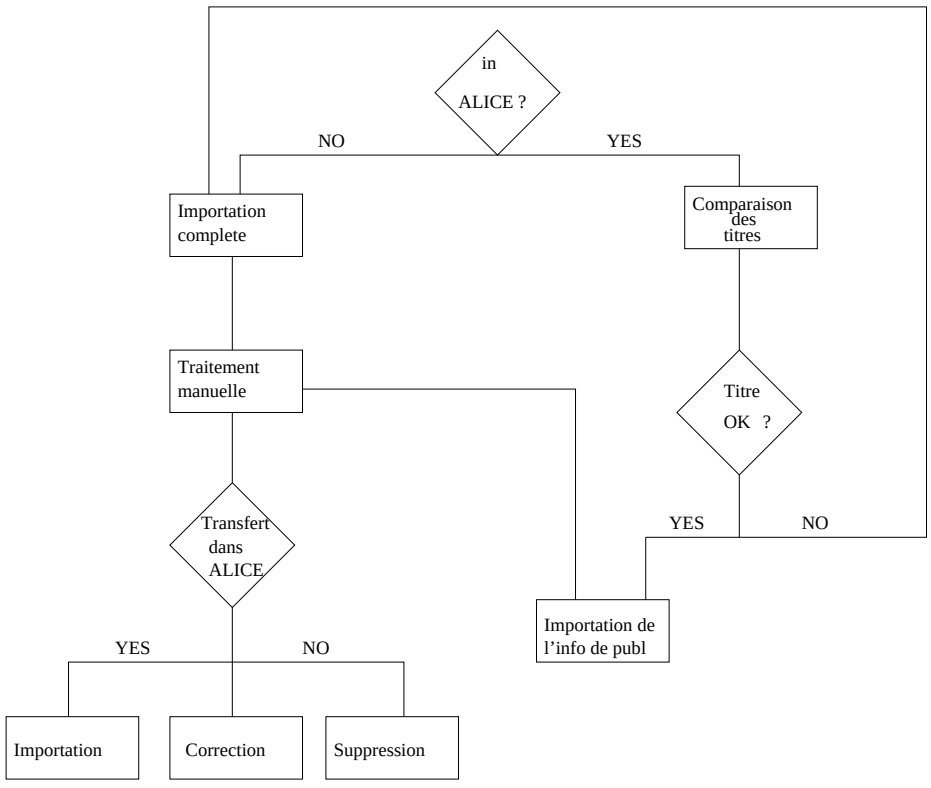

Fig. 4 : Algorithme du programme resultat.sh

Si le programme détecte l'article dans ALICE alors il présente le titre trouvé dans INSPEC et celui trouvé dans ALICE.
C'est à la documentaliste de la section qui sera chargée du programme, Catherine Cart, de décider si les articles sont les mêmes ou non.

S'ils sont identiques, il faut juste vérifier et importer l'information de publication si celle-ci n'existe pas ; sinon, l'article n'est pas dans la base et il faut alors l'importer entièrement.
Il est aussi possible de le supprimer ou de le corriger immédiatement par l'intermédiaire de *emacs*, un éditeur de texte souvent utilisé sous Unix.

Copie d'écran dans le cas ou l'article est déjà dans ALICE

```
************* Record dans ALICE ***************

Titre INSPEC: Pre-thermalization dynamics: initial conditions
for QGP at the LHC and RHIC from perturbative QCD
System Number INSPEC: 5951480
Annee de publication: 1997
Titre du periodique: Progress of Theoretical Physics Supplement
Volume: no.129
Page: 1-10
-----------------------------------------------
Titre ALICE:  Pre-thermalization dynamics: initial conditions
for QGP at the LHC and RHIC from perturbative QCD
System Number ALICE:  0256293
LKR:  . <I>Subm. to:</I>International School on the Physics of
Quark-gluon Plasma Hiroshima, Japan  ; 3 - 6 Jun 1997  .
 <I>Publ. in: </I>   Proceedings   Prog. Theor. Phys., Suppl.  0
***********************************************
Resultat de la comparaison entre le titre INSPEC et le titre Alice

 Pour mettre a jour l'information de publication : Taper 1
 Pour importer la totalite du record : Taper 2
 Pour supprimer le record : Taper 3

1

********* Information de publication **********
CER 0256293 PR    L $$pProgress of Theoretical Physics
Supplement : 2 no.129 (1997)$$c1-10
CER 0256293 YR    L 1997
CER 0256293 SW    L $$sn $$w9735 $$ya98
CER 0256293 BA    L 13
***********************************************

 Les champs sont corrects, pour importation: Taper 1
 Pour correction: Taper 2
 Pour suppression: Taper 3
```

Lorsque le programme ne trouve pas l'article dans la base CERN, il crée automatiquement la notice bibliographique ALICE et la présente tel qu'elle sera importée.
Catherine Cart doit vérifier et corriger d'eventuelles erreurs. Si elle juge que la structure convient alors la notice peut être importée. Sinon, il est toujours possible de la corriger ou de la supprimer.

Copie d'ecran dans le cas ou l'article n'est pas dans ALICE

```
************** New ALICE record ***************
CER 3112230 AU2   L Chliapnikov, P V
CER 3112230 AU    L Uvarov, V A
CER 3112230 OS    L $$i5951476
CER 3112230 PR    L $$pPhysics Letters B : 423, no.3-4 (1998)$$c401-6
CER 3112230 AF    L Inst. of High Energy Phys., Protvino, Russia
CER 3112230 LN    L eng
CER 3112230 NI    L CC9808
CER 3112230 YR    L 1998
CER 3112230 SW    L $$sn $$w9862 $$ya98
CER 3112230 TI    L Large excess of orbitally excited mesons at LEP
CER 3112230 BA    L 13
***********************************************

 Les champs sont corrects, pour importation: Taper 1
 Pour correction: Taper 2
 Pour suppression: Taper 3
```

Toutes les informations extraites sont donc vérifiées avant importation dans la base CERN. En effet, il serait dangereux d'importer automatiquement et sans contrôle des notices dans la base. L'importation des données reste donc semi-automatique.

2.5 Application de cette procédure automatisée

2.5.1 Résultats

Cette procédure sera à appliquer chaque mois et correspond à environ deux jours de travail. J'ai réalisé un 'mode d'emploi'[12] qui résume la méthode à employer.

Le programme a été testé au fur et à mesure de son développement sur des fichiers d'entrée contenant peu de notices. Et les corrections du programme ont été faites en conséquence.

Puis, nous l'avons testé sur un fichier plus volumineux.

- Les premières statistiques sont très encourageantes pour la poursuite de ce projet. L'interrogation de INSPEC sur les mises à jour des mois de mai à juin 98 nous a donné environ 800 références.
  Sur ces 800, 250 notices bibliographiques ont été importées entièrement et 215 ont été mises à jour. Les autres références étaient soit déjà présentes dans le catalogue du CERN soit ne correspondaient pas á l'activité du CERN.

- Le problème des doublons ou d'articles différents qui auraient le même titre n'est pas important puisqu'avant la publication du rapport annuel, une liste des titres avec le `SYSNO` est établie. Ainsi, il est possible de supprimer tous ces doublons.

- Nous nous sommes aussi aperçues que certains articles non trouvés dans la base du CERN font partie de conférence qui ont été dépouillées manuellement mais n'ont pas été détectés. Cette procédure permet donc de déceler des *By-Passed* oubliés.

- La première partie du programme est assez lente. En effet, le programme lit chaque ligne du fichier obtenu par DIALOG et ce fichier est assez volumineux. La lenteur du programme est aussi due au langage utilisé, il aurait été plus rapide en C.

- Bien que la plupart du bruit ait été éliminé, il reste quand même des articles extérieurs à l'activité du CERN comme, par exemple les ouragans dont les noms correspondent à des expériences du CERN et des atlas en cardiologie, Atlas étant aussi une expérience du CERN.

[12]Mode d'emploi de la procédure, cf. annexe L

2.5.2 Optimisations possibles du programme

Des modifications ont été et pourront être apportées par la suite à ce programme.

- Il sera possible d'importer d'autres documents que les *By-Passed*. Il suffit de modifier le fichier de départ, le programme reste valable pour n'importe quelle notice bibliographique provenant d'INSPEC.
- Jean-Yves Le Meur a ajouté un programme déjà existant qui corrige automatiquement les abréviations des titres des périodiques ce qui évite d'effectuer un lourd travail manuel de correction.
- Un lien hypertexte vers le résumé de INSPEC pourra aussi être envisagé.

Chapitre 3

Autres activités

3.1 Détection des erreurs dans la liste hebdomadaire des pré-tirages

Toutes les semaines, une liste des derniers pré-tirages entrés dans la base du CERN est publiée. Catherine Cart, qui est responsable de cette liste, effectue chaque semaine des corrections sur environ 500 notices bibliographiques. J'ai réalisé un document[1] qui résume la méthode utilisée par Catherine afin d'automatiser par la suite la correction de certaines erreurs. Cela m'a permis de comprendre le catalogage dans différentes sous-bases.

La correction des erreurs de numéro de sous-bases, de sous-champs s'effectue à l'aide d'une fonction d'ALEPH, `limit`. C'est un filtre qui sélectionne certaines notices en fonction des champs et sous-champs.

Exemples de `limit`

- Les champs *Sujet* sont obligatoires
  Pour vérifier, on effectue une limite :
  `lim total r su a z`
  Cela correspond à la sélection sur le nombre total de notices des champs `SU` qui ne sont pas vides.
  Si le nombre trouvé diffère du nombre total, certaines notices n'ont pas le champs `SU` rempli et il faut les corriger.

- Vérification au niveau des conférences.
  Tous les pré-tirages de la semaine appartenant à la base 12 (base

[1] Détection des erreurs de la liste hebdomadaire des *preprints*, cf. annexe M

des pré-tirages soumis à une conférence) doivent posséder un numéro système (sous-champs b) et un code conférence (sous-champs d).
On effectue alors les limites suivantes :

```
lim total r lkr b 0 9
lim total r lkr d a z
lim total s ba 12
```

Ces trois limites doivent donner les mêmes résultats. S'ils diffèrent, c'est qu'il y a des erreurs et qu'il faut corriger.

3.2 Programme de comparaison des listes d'auteurs

Certains pré-tirages sont écrits par des équipes de chercheurs très nombreuses (les collaborations) qui contiennent parfois plus de 500 auteurs. Jocelyne Jerdedet vérifie manuellement pour chacun de ses articles que tous les auteurs de la collaboration sont bien cités et bien orthographiés (accents, translittérations, particules).

J'ai établi un programme[2] qui permet de formater les deux fichiers d'auteurs sous une même forme puis de comparer chaque auteur grâce à la commande `diff`.
Le fichier résultat de ce programme fait apparaitre les différences entre les deux fichiers initiaux et donne ainsi la liste des auteurs manquants ou mal orthographiés.
Environ 5 pré-tirages de plus de 200 auteurs sont vérifiés ainsi chaque semaine.

3.3 Statistiques sur l'interrogation du catalogue du CERN

Chaque fois qu'une personne interroge le catalogue de la bibliothèque du CERN par l'intermédiaire du WEB, un `.log` est crée. Il contient le nom de la machine qui interroge, le jour, la date, l'heure et la requête.
Chaque jour, un fichier contenant tous les `.log` de la journée est crée.

[2]cf. annexe N

Exemple de .log

```
rsplus10.cern.ch - - [23/Jul/1998:00:27:09 -0100] "GET /search/
adsearch?searchtype=adsearch&uid=system_number&base=Preprints&
freetext1=wau%3DFrittelli+and+wti%3DNote+on+the+propagation+of
%23&searchrange=20&field1=all HTTP/1.0" 200 1863
```

Jean-Yves Le Meur m'a proposé de réaliser quelques statistiques sur les requêtes effectuées sur le catalogue WEB du CERN à l'aide d'un logiciel *Getstat*. *Getstat* analyse automatiquement le fichier contenant tous `.log` et produit un 'rapport' contenant le nombre de requêtes html, le nombre d'interrogation par suffixe, par domaine...
J'ai du réaliser des scripts Shell qui permettent de sélectionner des `.log` particuliers : par exemple, sélectionner les requêtes portant sur l'interrogation de la base *preprint*. *Getstat* produit ensuite les résultats portant sur ce fichier spécifique.
C'est le travail que je réalise actuellement et je n'ai pas encore de résultats d'analyse à ce jour.

3.4 Divers

- J'ai aussi eu l'occasion de participer à des *meetings* concernant l'évolution de la base de donnée CERN. En effet, la bibliothèque va bientôt changer la version de ALEPH pour passer à ALEPH 500. Cette version utilise le format US Mark et la base CERN va donc devoir convertir ses champs dans ce format.

- J'ai travaillé essentiellement sur des stations UNIX ce qui m'a permis de me familiariser avec cet environnement.
  De plus, pour le rédaction de ce rapport, j'ai utilisé le logiciel LaTeXqui est un formateur de texte très souvent utilisé dans la rédaction de rapports scientifiques [13,14]. Le logiciel `xfig` m'a permis de réaliser les différents schémas.

Conclusion

Ce stage de fin d'études a été riche en apprentissage. En effet, j'ai découvert le milieu de la bibliothèque et ses différentes facettes. Je réalise maintenant tout le travail qui se cache derrière l'interface WEB du catalogue de la bibliothèque du CERN `http: //wwwas.cern.ch/library/`.

Mon projet a été très formateur puisque j'ai programmé dans un langage que je ne connaissais pas et les difficultés rencontrées m'ont permis d'apprendre beaucoup sur moi-même et sur l'aspect programmation.
L'automatisation de tâche manuelle est un sujet très intéressant puisqu'il faut proposer une solution et la mener à terme. Cela permet de se poser de nombreuses questions notament sur les méthodes de travail.

Ce projet reflète la volonté du service d'information scientifique du CERN d'automatiser les tâches manuelles. D'autres projets menés parallèlement proposent des importations automatiques dans la base de données du CERN comme par exemple, des importations d'information de publication de la base de données UNCOVER.

Une poursuite de l'automatisation d'une partie du traitement de la littérature grise pourra être envisagée en effectuant la démarche inverse de ce programme : il serait possible de comparer les pré-tirages avec la base de données INSPEC. Si ceux-ci sont détectés dans INSPEC, c'est qu'ils ont été publiés : il faut donc importer l'information de publication.

Bibliographie

[1] HERMAN, Armin ; KRIGE, John ; MERSITS, Ulrike ; PESTRE, Dominique. *History of CERN*. Amsterdam : Elsevier Science Publishers B.V., 1987. 600 p. ISBN 0-444-87037-7

[2] *CERN Rapport Annuel 97, vol.1*. Genève : CERN, 1997, 48 p.

[3] *CERN Rapport Annuel 97, vol.3*. Genève : CERN, 1997, 219 p. ISSN 0304-2901

[4] *Aleph Manual Version 3.2.5, vol. 1*. Jérusalem : Ex Libris, 1995.

[5] *Aleph Manual Version 3.2.5, vol. 2*. Jérusalem : Ex Libris, 1995.

[6] *INSPEC User Manual*. London : The Institution of Electrical Engineers, 1993. ISBN 0-85296-494-3

[7] Institute of Electrical Engineeer. (Page consultée le 7 septembre 1998). *INSPEC Information*. Adresse URL :

`http://www.iee.org.uk/publish/inspec/`

[8] RIFFLET, Jean-Marie. *La programmation sous UNIX*. Paris : Ediscience, 1993. 630 p. ISBN 2-9105525-013-3

[9] ROCHKING, Marc. *Advanced Unix Programming*. Englewood : Prentice-Hall, 1985. 265 p. ISBN 0-13-011800-1

[10] Boston University. (Page consultée le 7 septembre 1998). *What does script do*. Adresse URL :

`http://cs-www.bu.edu/help/unix/what_does_script_do_.html`

[11] MAIRE Gilles. (Page consultée le 7 septembre 1998). *Un nouveau guide Internet*. Adresse URL :

http://www.cur-archamps.fr/mirror/ungi/

[12] University of Kansas. (Page consultée le 7 septembre 1998). *Lynx users guide version 2.3*. Adresse URL :

http://www.cc.ukans.edu/lynx_help/Lynx_users_guide.html

[13] GOOSSENS, Michel ; MITTELBACH, Frank ; SAMARIN, Alexander. *The LaTeXcompanion*. New York : Addison-Wesley, 1994. 528 p. ISBN 0-201-54199-8

[14] ROLLAND, Christian. *LaTeXguide pratique*. Paris : Addison-Wesley France, 1995. 345 p. ISBN 2-87908-104-1

Les règles bibliographiques s'inspirent de la norme AFNOR Z44-005.

Annexe A

Exemples de document de la base ALICE

Annexe B

Circulaire 29

Annexe C

Statistiques

Annexe D

Rapport préliminaire
Automatisation du traitement du rapport annuel

Annexe E

Extrait de l'index du rapport annuel 97 vol.III

Annexe F

Programme *main.sh*

Annexe G

Programme *record.sh*

Annexe H

Extrait du fichier obtenu par DIALOG

Annexe I

Programme *extract.sh*

Annexe J

Programme *cleanrecord.sh*

Annexe K

Programme *resultat.sh*

Annexe L

Procédure d'utilisation du programme

Annexe M

Détection des erreurs sur la liste hebdomadaire des *preprints*

Annexe N

Programme de comparaison des auteurs