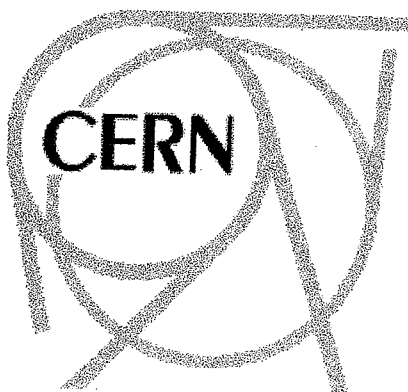




Année 2001-2002

**Contribution à l'automatisation partielle du
traitement de la littérature grise au Service
d'Information Scientifique du CERN.**

De l'analyse des sources à l'importation des documents.

Mémoire de stage sous la direction de
Monsieur Alain Lelu
de Madame Ingrid Geretschläger
et de Madame Jocelyne Jerdelet

Présenté par Anne Gentil-Beccot
DESS Système d'information Documentaire en Ligne
Université de Franche-Comté, Besançon.

CONTRIBUTION A L'AUTOMATISATION PARTIELLE DU
TRAITEMENT DE LA LITTÉRATURE GRISE AU SERVICE
D'INFORMATION SCIENTIFIQUE DU CERN.
De l'analyse des sources à l'importation des documents.

Anne GENTIL-BECCOT

Sous la direction de
Monsieur Alain LELU
Madame Ingrid GERETSCHLÄGER
et de Madame Jocelyne JERDELET

Résumé :

Ce travail présente la mission que j'ai eu à accomplir pendant mon stage de fin d'étude au Service d'Information Scientifique du CERN. Il s'agissait de participer à l'activité d'importation automatique de documents de littérature grise au sein de la base locale. Ce processus est possible grâce au programme de transformation de données *Uploader*. Ce mémoire tente de montrer quelle a été la méthode suivie : l'analyse des sources à exploiter et l'importation des données.

Mots-clefs :

Littérature grise – Prépublication - Importation automatique – Transformation de données – Uploader .

Abstract :

This work presents my activities during my training period in the Scientific Information Service at CERN. I worked on the automatic method of importation of grey literature's meta data with a program called Uploader. This report explains what was my method: the database sources's analysis and the data's importation.

Keywords :

Grey literature – preprints – automatic importation – Meta data's transformation – Uploader.

Remerciements

Je tiens à remercier tout particulièrement Corrado Pettenatti, chef de groupe du Service d'Information Scientifique et Ingrid Geretschlager, responsable de la section de la Gestion des Documents pour m'avoir permis d'effectuer mon stage au sein de leur équipe.

Je voudrais exprimer ma reconnaissance à Monsieur Alain Lelu, professeur à l'université de Franche-Comté pour avoir accepté d'être mon tuteur de stage.

Je remercie tout aussi vivement Jocelyne Jerdelet, responsable de l'unité des prépublications, pour son aide précieuse et ses conseils pendant mon travail, ainsi que toute l'équipe du support informatique pour leur disponibilité.

Merci, également, à toute l'équipe du SIS qui m'a accueillie si chaleureusement pendant ces quatre mois, et un merci tout particulier à Tullio Basaglia qui m'a supportée dans son bureau pendant tout ce temps...

Je tiens, enfin, à exprimer ma reconnaissance à tous ceux qui m'ont apporté leur soutien pendant la rédaction de ce mémoire.

Avertissement au lecteur

Etant donné que le cadre du stage est un organisme international, de nombreux termes en rapport avec le travail effectué sont en anglais. Parfois, il est plus pertinent de les laisser dans cette langue. Cependant, tous les mots de langue anglaise sont inscrits en *italique*.

Sommaire

Introduction	6
I. Présentation du cadre du stage et de la mission à accomplir.....	7
1. Le CERN	7
2. Le Service d'Information Scientifique du CERN	8
a. Les objectifs du SIS	8
b. L'organisation	9
c. Les collections	9
d. Le logiciel de gestion documentaire utilisé par le SIS	10
3. Définition de ma mission de stage.	12
a. La Section de Gestion des documents	12
b. L'Uploader	13
c. La mission	15
II. Réalisation de la mission : analyse des sources et importation des données.....	16
1. Etat des lieux	16
2. Liste des sources explorées	18
3. La phase d'analyse	19
a. Analyse informationnelle	20
b. Analyse documentaire	21
c. Analyse matérielle : Accès direct aux informations.....	23
4. Configuration du programme Uploader	25
a. La séparation des données.....	25
b. Les trois fichiers de configuration.....	26
c. Upload, Correct, Append.....	28
d. Les "knowledge bases"	29
e. Le texte intégral.....	30
5. Importation et validation des informations importées.....	30
6. Bilan des importations.....	32
III. Bilan et perspective pour cette forme de traitement automatique du document.	34
1. Intérêts et limites du programme Uploader.....	34
a. Intérêts	34
b. Les limites	36

2. Un regard sur les sources exploitées.....	37
a. Il existe des sources non adaptées à l'Uploader.....	38
b. Un manque de stabilité de certaines sources.....	38
c. Une redondance des données.....	39
d. Quel bilan peut-on conclure de tout cela ?.....	39
3. Uploader : combler un manque en attendant mieux ?.....	40
4. Une remise en cause du système traditionnel.....	41
Conclusion.....	43
Bibliographie.....	44
Annexes.....	
ANNEXE 1.....	46
ANNEXE 2.....	47
ANNEXE 3.....	49
ANNEXE 4.....	51
ANNEXE 5.....	56
ANNEXE 6.....	60
ANNEXE 7.....	68
ANNEXE 8.....	70
ANNEXE 9.....	72
ANNEXE 10.....	74
ANNEXE 11.....	77
	79

Introduction

Le CERN est le plus grand centre mondial de recherche en physique des particules, il accueille quelque 6500 chercheurs scientifiques de plus de 80 nationalités différentes. Autant dire que la quantité de documents créés ou simplement exploités par cet organisme est très importante. On comprend alors le rôle que peut prendre son Service d'Information Scientifique. C'est là que j'ai été amenée à effectuer mon stage de fin d'étude de quatre mois.

J'ai été affectée à la section Gestion des documents et plus particulièrement, dans l'unité des prépublications.

Les prépublications, *preprints* en anglais, jouent un rôle de plus en plus important dans la communication scientifique. Les *preprints* sont les articles qui n'ont pas encore été publiés dans un périodique mais qui sont tout de même diffusés pour permettre aux chercheurs d'accéder plus rapidement à l'information. Une publication peut prendre, en effet, plusieurs mois, voire même deux ans, le *preprint* lui peut être diffusé instantanément par le chercheur (sur des serveurs spécialisés par exemple). Or le SIS du CERN se doit de fournir, le plus rapidement possible, le maximum d'informations à ses chercheurs, voilà pourquoi, il favorise au maximum l'acquisition et la diffusion des *preprints*.

Pour mener à bien ce travail, l'unité des prépublications a mis en place, depuis quelques années un processus d'importation automatique des documents, grâce à l'utilisation d'un programme de transformation de données nommé *Uploader*, créé par le groupe du support informatique de la bibliothèque. Ce procédé était à l'origine destiné à rapatrier dans la base locale les documents des chercheurs du laboratoire qui omettaient de soumettre leurs travaux directement au CERN. Aujourd'hui l'objectif est plus important, on importe, grâce à ce processus, des documents qui proviennent du monde entier ; la politique d'acquisition est donc modifiée. Ce programme de transformation de données a permis d'augmenter considérablement le nombre de notices importées.

J'ai eu pour mission d'utiliser ce processus d'importation pour exploiter de nouvelles sources documentaires et d'en importer les données afin d'enrichir la base de données du SIS. Pour mener à bien cette mission, il m'a fallu accomplir différentes tâches :

- L'analyse de sources documentaires
- L'exploitation de ces sources
- L'utilisation du programme *Uploader*
- L'intégration des sources dans la base.

Ce stage a été pour moi une expérience très enrichissante, car j'ai pu aborder différents problèmes et j'ai pu m'initier à cette forme d'importation automatique utilisée au CERN. Elle est très intéressante car elle ouvre des perspectives nouvelles dans le rôle du documentaliste. En effet, le but de ce processus est enrichir la base locale de documents utiles aux chercheurs du laboratoire et ainsi d'éviter qu'ils aient plusieurs recherches à faire dans différentes bases pour trouver ce dont ils ont besoin. Le documentaliste dans ce travail prend en charge une partie de la recherche et devient donc l'intermédiaire entre l'information et l'utilisateur.

Dans un premier temps, je vais présenter quel a été le cadre de mon stage. Je m'arrêterai ensuite sur l'accomplissement de la mission en elle-même, en expliquant la méthode de travail suivie. Enfin, je tenterai de mener une réflexion sur le bilan de mon stage et, plus généralement, sur cette forme de traitement de documents.

I. Présentation du cadre du stage et de la mission à accomplir

1. Le CERN

Situé aux abords de Genève, à la frontière entre la France et la Suisse, le CERN est l'un des plus grands centres de recherche du monde dans le domaine de la physique des particules. En 1954, lorsque le laboratoire a été créé, l'acronyme CERN signifiait Conseil Européen pour la Recherche Nucléaire ; aujourd'hui il se nomme l'Organisation Européenne pour la Recherche Nucléaire. Le nom a changé à plusieurs reprises, l'acronyme, lui, est resté.

C'est après la deuxième guerre mondiale, suite à l'exode massif de la communauté scientifique aux Etats-Unis, que certains pays d'Europe ont pris conscience qu'en unissant leurs forces financières et scientifiques pour créer un laboratoire de recherches commun, ils seraient capables de concurrencer les Etats-Unis.

En 1954, 12 états membres ont signé sa convention constitutive, aujourd'hui, il y en a 20. Ces pays financent le CERN proportionnellement à leur PIB. Il y a aussi 35 états non membres qui utilisent les structures du laboratoire pour leurs recherches, ils s'autofinancent. On trouve enfin les états observateurs.

Depuis, cinq accélérateurs ont été construits pour tenter de résoudre de nombreuses interrogations ; les physiciens étudient les particules fondamentales qui composent la matière et les forces qui unissent ces particules. Les objectifs de ces recherches ne sont ni militaires ni commerciaux.

Le CERN joue un rôle fondamental dans le développement des nouvelles technologies. En effet, pour leurs diverses expériences, les physiciens ont besoin de machines sophistiquées. Leurs recherches entraînent, par conséquent, la progression des connaissances dans d'autres domaines. Un des exemples les plus pertinents est l'invention du WEB, c'est au CERN qu'il a été créé pour améliorer la communication entre les scientifiques du monde entier. Il va sans dire que la recherche en physique des particules a des applications pratiques. On peut aussi citer les progrès médicaux liés à la thérapie à hadrons (les rayons dans le traitement contre le cancer) ou à l'utilisation du scanner médical...

Un tel laboratoire de recherche génère un échange d'informations considérable : les chercheurs ont besoin de documentations pour leurs travaux, de même, ils se doivent de diffuser les résultats de leur propre recherche. C'est pour cela que le Service d'Information Scientifique prend une place très importante au CERN. En effectuant un travail d'acquisition de documents scientifiques nécessaires aux activités du CERN, et en stockant et diffusant les publications du CERN, il fait en sorte que le flux d'informations circule du CERN au reste du monde et réciproquement.

2. Le Service d'Information Scientifique du CERN

Ce service est né en même temps que le CERN en 1954. Il est situé au sein de la division ETT, *Education and Technology Transfert*¹. Le CERN est, en effet, composé de divisions, chaque division ayant son activité propre. L'ETT est responsable de la communication des résultats scientifiques réalisés au CERN en collaboration avec les différents instituts.

Le Service d'Information Scientifique est composé d'une bibliothèque centrale ainsi que de quatre bibliothèques satellites et d'une bibliothèque de service légal.

a. Les objectifs du SIS

- Il a pour mission principale de gérer la bibliothèque. Par conséquent, il doit acquérir et gérer des informations dans tous les domaines pertinents pour le CERN et rendre celles-ci accessibles de façon simple et pratique à la communauté de la physique des particules présente au CERN, mais aussi à la communauté scientifique mondiale.
- Il se doit aussi de gérer les Archives Historiques et Scientifiques du CERN (*Historical and Scientific Archives*).
- Il assure la distribution des publications du CERN : les rapports, les *preprints* soumis par les chercheurs.
- Le SIS participe activement au développement des nouvelles techniques de documentation, il fait partie de AILIS (*Association of International Librarians and Information Specialists*)². On peut noter aussi sa participation à l'Open Archive Initiative³.

Le SIS du CERN a donc une double mission, dans un premier temps, il se doit d'apporter aux chercheurs du laboratoire la documentation nécessaire pour la recherche, dans un second temps, il lui faut rendre disponible le travail effectué au CERN pour le reste de la communauté scientifique mondiale. Il a un double rôle d'acquisition et de diffusion.

¹ Voir l'organigramme de la division en annexe 1

² Association internationale des bibliothécaires et des spécialistes de l'information.

³ L'objectif de l'Open Archive Initiative est de favoriser la diffusion de l'information scientifique et technique (en particulier les preprints). Pour cela, différentes bibliothèques et centres de documentation tentent d'adopter des normes communes afin d'échanger facilement des données sans de lourdes modifications.

Lorsque les utilisateurs se rendent à la bibliothèque du CERN, ils sont frappés par la performance du service de prêt interbibliothèque. Il est vrai que cette activité est importante au CERN, mais elle ne représente qu'une partie seulement du travail du SIS, les utilisateurs ne réalisent pas forcément l'ampleur du travail qui est fait en amont de leur utilisation.

Le SIS bénéficie, en effet, de moyens technologiques considérables qui lui permettent de fournir un travail de qualité pour servir ses lecteurs.

Le CERN consacre au travail de gestion et de conservation du savoir effectué au SIS un budget qui s'élève à 1 250 000 CHF, cela représente 0,1% seulement de son budget. C'est très peu comparativement à l'importance du travail du Service de l'Information Scientifique pour les activités de recherches au sein du CERN et dans le monde entier.

b. L'organisation

Le SIS comprend quatre sections :

- Section Gestion des documents¹ : cette section est divisée en deux unités, l'unité des livres (gestion des acquisitions de monographies) et l'unité des prépublications² (lieu du stage) .
- Section Gestion des périodiques : ce service s'occupe des revues papier et électroniques. Désormais, il s'oriente résolument vers une politique «tout électronique».
- Section des utilisateurs : elle gère les prêts entre bibliothèques, la construction des pages Web du site de la bibliothèque, la distribution papier des publications CERN ; on s'occupe ici de tout ce qui concerne directement l'utilisateur final.
- Section des Archives Historiques et Scientifiques : créée en 1980, cette section gère le fonds ancien du CERN.

c. Les collections

En ce qui concerne les collections, on peut dire que deux principaux types de documents sont proposés par le SIS : les documents sur support papier classés au sein de la bibliothèque, et les documents électroniques.

¹ Document Management Section en anglais.

² Les prétrages ou preprints 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 prépublication et la publication.

La composition des collections:

- **Des monographies**

Le SIS enregistre 3 000 titres de livres publiés par an.

- **Des périodiques**

On compte 600 journaux scientifiques en version imprimée, il y a aussi les collections de périodiques qui ont cessé de paraître et ceux auxquels la bibliothèque n'est désormais plus abonnée.

On a, enfin, 1100 titres en version électronique. La section des périodiques privilégie de plus en plus ce support, on se dirige vers une politique du tout électronique.

- **Des conférences**

Il y a actuellement 40 000 titres et comptes rendus de conférences (*proceedings* en anglais), dont environ 800 nouveaux par an.

- **De la littérature grise : *preprints* et rapports**

320 000 documents sont disponibles, dont seulement 8 % proviennent du CERN. Ces documents sont accessibles en microfiches (environ 15 000), en version papier ou en version électronique (depuis 1994, les documents sont disponibles uniquement en version électronique, sous différents formats, seuls les documents CERN sont conservés sous forme papier pour l'archivage).

Depuis l'année 200, plus de 50 000 *pretrages* (*preprints* en anglais), provenant de plus de 100 sources différentes, sont versés chaque année dans la base de données, avec un accès au texte intégral.

d. Le logiciel de gestion documentaire utilisé par le SIS

Le logiciel utilisé à la bibliothèque du CERN est ALEPH300. ALEPH (*Automated Library Expandable Program*) est un logiciel de gestion de données développé au sein de l'Hebrew University à Jérusalem en Israël. Il est produit par Ex-Libris, société israélienne spécialisée dans le domaine de l'informatisation de médiathèques. La version initiale a été créée il y a 20 ans. Depuis, plusieurs versions ont été mises au point, conduites par les besoins des bibliothèques du monde entier, la dernière en date est ALEPH 500.

Le service d'information scientifique du CERN utilise actuellement la version 300. Le passage à la nouvelle version (ALEPH 500) est prévu, mais aucune date n'est actuellement fixée. Il faut dire qu'Aleph 300 a été configurée pour son utilisation spécifique au CERN et que ce changement de version risque d'entraîner divers problèmes passagers.

Les grands centres de documentations n'utilisent généralement pas ALEPH car ce gestionnaire de bases de données ne possède pas de véritable thesaurus et les modules offerts ne sont pas pratiques pour la gestion courante (commandes, budgets, prêts...). Pourtant, il permet une gestion très flexible et peut être modifié en fonction des besoins de l'utilisateur : le nombre de champs pour un enregistrement, ainsi que leur longueur ne sont pas limités, le format d'entrée est défini par l'utilisateur.

Ce système est utilisé par 3 millions de personnes sur 560 sites dans 44 pays. Le système Aleph a été choisi en France et en Suisse par de nombreux établissements prestigieux. Nous pourrions citer à titre d'exemple les écoles Polytechniques de France et de Suisse ainsi que nombreux établissements universitaires du monde entier (Suisse, Autriche, Espagne...).

Il peut s'adapter au langage de chaque bibliothèque, il offre 20 interfaces différentes de langue, reconnaît de nombreux caractères (alphabet hébreu, grec, arabe, latin, cyrillique et grec...). En cela, il est très utile pour les documents contenant des formules mathématiques traités au CERN.

Il possède les modules classiques : la recherche documentaire, le catalogage, le prêt entre différentes bibliothèques, la gestion des fournisseurs, des commandes, des prêts, des lecteurs, des périodiques...

La version 300 d'ALEPH fonctionne principalement sous UNIX alors que l'interface de la version récente ALEPH 500, est de type Windows. Cela implique l'utilisation du langage LaTeX¹ 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 car modifiée et améliorée par les informaticiens du support informatique de la bibliothèque du CERN selon ses besoins spécifiques.

La base de données globale du SIS est constituée de plusieurs sous-bases qui sont indépendantes ; elles ont chacune une structure adaptée au type de documents qu'elles contiennent. Par exemple, les *preprints* font partie de la base PREPS et les articles publiés appartiennent à la base ANALP. Il serait un peu fastidieux de toutes les énumérer. Dans le travail que j'ai eu à faire, je travaillais surtout sur ces deux bases.

La recherche sous ALEPH 300 est extrêmement pertinente, elle peut s'effectuer depuis deux interfaces : l'interface Web (pour les utilisateurs) et l'interface UNIX en utilisant le protocole TELNET (pour les documentalistes).

Concernant l'interface Web, il existe un mode d'interrogation où l'utilisateur peut effectuer une recherche simple dans tous les champs (titre, auteur, résumé, numéro de rapport...), mais aussi combiner les champs de recherche en utilisant les opérateurs booléens. Il peut aussi faire une recherche pour trouver un document (principalement les livres) en utilisant la cote CDU (Classification Décimale Universelle).

Pour ce qui est de l'interface Telnet, il y a deux modes d'interrogation :

- La recherche par index : on utilise la fonction b pour « Browse » et on choisit un champ d'entrée, par exemple le champ de l'auteur (AU). Ainsi, il faut donc taper B AU= Nom de l'auteur
- La recherche par mots permet de trouver un titre précis dont on est parfaitement sûr. Pour cela, on utilise la fonction f pour « Find ». Par exemple :
« F mots du titres »
ou « F nom d'auteur ? » (Troncature)

¹ Formateur de texte utilisé pour la rédaction de formules mathématiques

On peut, en outre, croiser les éléments de recherche, et ainsi construire une équation avec des opérateurs booléens. Remarquons un petit plus avec la recherche par l'interface TELNET : le personnel de la bibliothèque peut effectuer des recherches en utilisant la fonction *limit searches* (recherches limitées) qui propose des filtres sur les champs non indexés.

Attardons-nous un moment sur l'interface Web. Le portail Weblib utilisé au CERN est un point d'accès mondial à l'information et à la documentation dans le domaine de la physique des hautes énergies. Le but de ce portail est de proposer un accès à des ressources scientifiques pertinentes et à des bases de données locales par une interface unique. Cette interface donne accès à divers types de documents (*preprints*, rapports, articles, livres, revues, normes, thèses, archives scientifiques, vidéos de conférences, annuaire des instituts de physique, photographies, coupures de presse...) et permet une interconnexion des informations. Ainsi, lors d'une recherche, on offre un accès à diverses sources, pages ou sites, en relation avec la recherche. Ce concept est connu sous diverses appellations. On peut citer par exemple SFX, CrossRef ou Go direct. Ce dernier a été élaboré par le Service d'Information Scientifique du CERN en 1997, puis proposé à l'*American Physic Society* qui l'a adopté.

3. Définition de ma mission de stage.

a. La Section de Gestion des documents

C'est au sein de cette section que j'ai évolué pendant quatre mois, et plus précisément dans l'unité des Prépublications. Comme il a déjà été dit plus haut, cette unité se charge de gérer tout ce qui concerne la littérature grise pour le SIS (Acquisition, mise à disposition des documents...).

Depuis quelques années les *preprints* ont pris une place très importante dans la diffusion de la littérature scientifique. Les *preprints* sont les papiers qui ne sont pas encore publiés et, par conséquent, non soumis à l'approbation d'un comité de lecture. Josette de La Vega définit le préirage ou *preprint* «comme un article soumis à publication à une revue, non encore publié et destiné à l'être, qui circule de façon informelle dans la communauté. Il fait partie de ce qui est défini plus largement sous le terme de « littérature grise »"¹

Etant donné que les délais de publication sont parfois très importants, il est intéressant, pour un chercheur, de diffuser son travail avant ; il va de soi que le plus important pour lui est de permettre une large diffusion de ses recherches le plus rapidement possible, même si cela n'implique pas de rentabilité financière. Le *preprint* permet, en effet, de poser une antériorité dans un système donné. En attendant que le texte soit publié, le chercheur peut montrer ses résultats de recherches, et éviter ainsi que quelqu'un d'autre se les approprie pendant le temps relativement long de la publication.

Voilà pourquoi le SIS du CERN favorise au maximum l'acquisition de ce type de document. Il se doit d'abord de récupérer tous les travaux faits au CERN, mais comme cela ne suffit pas, il acquiert aussi des *preprints* venus des laboratoires de recherche du monde entier pour permettre aux physiciens un accès rapide et unique à l'information. Un

¹ [4] VEGA (de la) Josette F., *La communication scientifique à l'épreuve de l'Internet – l'émergence d'un nouveau modèle*, Villeurbanne, Presses de l'ENSSIB, 2000, p. 125

programme a été créé par le support informatique lié à la bibliothèque pour permettre d'importer de façon semi-automatique des données bibliographiques d'autres sources d'information : il s'agit de l'*Uploader*. Ce programme permet de transformer les données de la base initiale dans le format voulu pour la base du CERN. Ce processus permet d'importer les données bibliographiques ainsi que les textes intégraux des documents

Grâce à ce système, le CERN n'importe pas seulement des *preprints*, il importe aussi des notices d'articles publiés, des *proceedings* (i.e. comptes-rendus de conférences), des thèses...

Le SIS a pour objectif d'importer le maximum de données bibliographiques ; plus la notice est complète, plus elle est intéressante. Le SIS donne la priorité aux notices qui possèdent un accès au texte intégral. C'est, en effet, une des données les plus importantes. Cet élément n'est pas toujours présent dans les notices, aussi nous ne pouvons pas le prendre comme critère principal d'importation. Nous approfondirons ce sujet plus loin.

Ces importations sont d'abord faites sous forme de tests. S'il s'avère que la démarche est positive, le CERN demande une autorisation au propriétaire de la source pour pouvoir importer ses données en toute légitimité. Le CERN ne peut, en effet, se permettre d'exploiter les sources sans permission. Le SIS doit indiquer dans les notices importées leur provenance. Il arrive aussi que les gestionnaires de la source dont on veut importer les données demandent l'autorisation d'effectuer la même opération sur la base de données du CERN. Ils sollicitent, eux-même la permission de pouvoir puiser dans les données bibliographiques du CERN ; souvent, ils ont leur propre système d'importation.

b. L'Uploader

Il y a quelques années, l'unité *Preprint* du SIS du centre de documentation recevait les *preprints* des autres centres de recherche grâce aux *mailing-lists* par courrier traditionnel. Les chercheurs du CERN étaient sensés envoyer leur travaux à la bibliothèque. Or, lorsque l'on a décidé d'analyser les données proposées par INSPEC¹, on s'est rendu compte que l'on y trouvait des papiers du CERN qui n'avaient jamais été soumis à la bibliothèque. Il fallait donc trouver un moyen de les récupérer. C'est ainsi qu'un premier programme a été créé pour permettre d'importer ces « *by-passed* » à partir d'INSPEC. Ce programme a été effectué en collaboration avec le support informatique lié à la bibliothèque. Il a permis d'automatiser une partie du rapport annuel (Rapport qui recense chaque année toutes les publications CERN). Cependant, ce programme était encore trop peu performant et présentait certains inconvénients techniques. Il fallait, par exemple, que l'ordinateur soit allumé pendant le traitement d'un fichier. On devait aussi vérifier les notices *non-confident* (non fiables) en même temps que le fichier tournait. Ainsi, un traitement d'une centaine de notices pouvait prendre deux jours de travail, ce qui est très lourd. C'est ainsi que l'idée a été lancée de faire un programme plus ouvert qui permettrait de transformer des données bibliographiques des bases de données pour les mettre au format voulu par ALEPH. C'est ainsi que Martin Vesely alors étudiant en informatique au CERN a créé l'*Uploader*. C'est un système qui permet la transformation de données bibliographiques de différentes sources dans un format accepté par la base de données locale. Il a pour but de favoriser l'importation de notices bibliographiques quelle

¹ INSPEC est une base de données bibliographiques produite par l'*Institution Of Electronical Engineers*, elle contient presque 7 000 000 de notices. Cette base dépouille la plupart des revues et compte-rendus de conférences anglophones en sciences.

que soit leur origine. Au début, il n'y avait qu'une source explorée : Fermilab¹. C'est au fur et à mesure de l'utilisation de l'*Uploader* que l'on a pu analyser d'autres sources. Ainsi l'*Uploader* s'est construit petit à petit ; selon les besoins définis par l'équipe de documentalistes de l'unité *preprint*, le support informatique développe de nouvelles commandes destinées à rendre l'outil toujours plus performant. Le SIS s'occupe de définir quelles sont les fonctions utiles, et le support informatique se charge de l'aspect programmation. C'est un brillant exemple de la nécessaire collaboration entre les services informatiques et les spécialistes de la documentation.

La procédure de l'*Uploader* se décompose en plusieurs étapes :

- Analyse des données bibliographiques de la source.
- Transformation de ces données dans le format adapté à la base locale.
- Importation dans la base locale.
- Vérification de ces données en les comparant aux informations déjà contenues dans la base ALEPH via une recherche portant sur certains champs de la notice.
- Nécessaire intervention manuelle pour les notices litigieuses (*Non-confident records*)
- Plusieurs possibilités d'importations :
 - Update (Mise à jour des notices déjà existantes des notices déjà existantes dans la base)
 - Upload (Crée de nouvelles notices)

Chaque nouvelle source doit être configurée afin que les données qui y sont contenues puissent être utilisées. Ce fichier de configuration établit la structure de la notice initiale, ainsi que les fonctions qui doivent lui être appliquées afin de la transcrire dans le format adéquat.

Il y a trois étapes au cours de ce processus :

- séparation des notices (toutes les notices étant contenues dans un fichier unique)
- extraction des différents champs de chaque notice
- codage des champs dans le format local.

Une fois que les données ont été transformées dans le format de la base locale, la base ALEPH est interrogée, on y détermine ainsi leur présence (ou non). Trois cas sont possibles :

- La notice n'est pas trouvée dans ALEPH, elle est considérée comme nouvelle (*New*), dans ce cas la notice entière est importée avec un nouveau numéro de système.
- Le système trouve une occurrence de la notice, elle est considérée comme vérifiée (*Matched*). Plusieurs cas sont possibles : quelques champs sont corrigés (c'est la fonction *correct*), des champs sont ajoutés à la notice existante (fonction *append*) ou la notice est ignorée.
- Le système trouve plus d'une occurrence de la notice, elle est alors litigieuse (*Not confident*). Il faut alors vérifier chaque notice pour la valider ou non, et si besoin, corriger les champs.

Ce programme ne présente pas une interface particulièrement conviviale, rappelons, qu'il fonctionne sous UNIX. Mais cela n'est pas un problème puisqu'il offre de nombreuses fonctions. On pourrait le comparer à un autre programme de transformation de données : F2F (*Format to format*). Celui-ci offrait une interface plus

¹ Fermi National Laboratory, Batavia, Illinois.

agréable, cependant toutes les options étaient cachées. Il était donc plus difficile à utiliser.

c. La mission

Il s'agissait d'explorer de nouvelles sources documentaires afin d'importer, grâce à l'*Uploader*, les méta données dans la base de données du SIS. Ce projet implique des tâches assez diverses qui se complètent pour obtenir le résultat voulu.

Par conséquent, ma mission a été composée de plusieurs travaux qui, finalement, avaient tous le même objectif : permettre l'acquisition de nouvelles données au sein de la base afin d'offrir à l'utilisateur une recherche plus performante. En effet, si on importe des documents contenus dans des sources extérieures, le travail de recherche est facilité, au lieu d'effectuer plusieurs interrogations, l'utilisateur n'en fait qu'une¹.

- Il s'agit tout d'abord d'un travail d'analyse de sources : pertinence des contenus pour le CERN, stabilité du support (Base de données ou Page Web), accessibilité de l'information...
- Les tâches sont ensuite plus techniques : pour pouvoir importer automatiquement les données, il est nécessaire de configurer l'*Uploader*, c'est à dire de créer trois fichiers pour permettre l'extraction et la transformation des données dans le format voulu. Le travail s'apparente ici davantage à de la programmation.
- Travail de validation des données : correction des données passées dans l'*Uploader* et passage dans la base ALEPH. Cette activité reste fastidieuse, elle est la preuve que ce procédé d'importation n'est encore pas entièrement automatisé. L'intervention intellectuelle et manuelle de l'homme est nécessaire.

Voilà les étapes essentielles de la mission que j'ai eu l'opportunité d'accomplir. Il est clair que je n'ai pas eu à élaborer une méthodologie générale sur ce projet puisque l'*Uploader* est utilisé depuis déjà deux ans au CERN, et que d'autres avant moi ont réfléchi aux manières de rendre son utilisation optimale. Il s'agissait d'avantage de poursuivre le travail commencé, d'analyser de nouvelles sources et d'en exploiter le contenu. Par ailleurs, mon travail a permis de mettre en lumière les éléments qui étaient à améliorer dans le programme. J'ai donc travaillé en collaboration avec le support informatique pour développer de nouvelles fonctions sur l'*Uploader*.

Voici donc en quelques mots, le contexte de mon stage. Passons maintenant à la description plus détaillée de la manière dont j'ai réalisé le travail qui m'a été confié.

¹ Voir le schéma en annexe 2.

II. Réalisation de la mission : analyse des sources et importation des données

Analyser de nouvelles sources afin d'en exploiter les données, voici en quelques mots ce que j'ai dû accomplir pendant les quatre mois de mon stage au CERN. Mission assez déconcertante au premier abord, mais qui se révèle être d'un grand intérêt, par les différents travaux qui lui sont associés et les sujets de réflexion dont elle est la source.

Pour montrer comment j'ai travaillé, j'ai décidé d'expliquer quelles ont été les tâches que j'ai été amenée à accomplir :

- Analyse de sources documentaires
- Exploitation de ces sources, recherche d'une utilisation optimale
- Configuration du programme *Uploader*
- Importation automatique
- Gestion des nouvelles entrées dans la base : correction et validation

Je vais tenter de décrire la méthode suivie pendant cette mission ; ces explications seront illustrées par des exemples concrets rencontrés lors de mes travaux. Je ne peux, en effet, faire une étude linéaire de ce que j'ai effectué pendant mon stage. J'ai travaillé sur environ une dizaine de sources et il serait fastidieux de décrire les étapes du travail sur chacune d'entre elles. Voilà pourquoi, une étude transversale me semble plus appropriée, en exploitant quelques exemples issus de chacune des sources sur lesquelles j'ai travaillé.

1. Etat des lieux

L'importation semi-automatique de données effectuée grâce à l'*Uploader* est une opération qui se fait depuis déjà deux ans. La démarche est donc définie, l'outil a beaucoup été testé, il a pris son rythme de croisière.

Beaucoup de sources ont été explorées depuis la conception de l'*Uploader*. Ce sont essentiellement les bases de données constituées par les grands laboratoires de recherches (Fermilab à Chicago, DESY en Allemagne, ou Los Alamos...). Il y a aussi des bases commerciales comme INSPEC¹.

On importe principalement des documents de littérature grise (*preprints*, thèses, rapports...), mais aussi des articles publiés (références de publication, *proceedings*...).

Le SIS a des objectifs très clairs. Dans un laboratoire de recherche comme le CERN, il ne peut se disperser dans ses projets. Le SIS existe pour aider les scientifiques dans leurs

¹ Voir les statistiques de Catherine Cart sur les différentes sources déjà exploitées en Annexe 11

travaux. Etant donné que le CERN est un laboratoire en physique des particules, le SIS se doit de diriger ses acquisitions dans ce domaine. On note toutefois que les activités connexes à la recherche fondamentale sont très nombreuses, et comme ces activités nécessitent aussi de la documentation, le SIS doit être en mesure de pouvoir la fournir. C'est pourquoi les acquisitions concernent des sujets assez nombreux dans le domaine des sciences. Voici la liste des sujets que l'on peut trouver indexés dans le catalogue du CERN¹. Cette liste est écrite en anglais sur l'interface de recherche, je la retranscris telle qu'elle est :

- *Particle physics*
- *Particle physics, experiment results*
- *Particle physics, phenomenology*
- *Particle physics, theory*
- *Particle physics, lattice*
- *Nuclear physics*
- *General relativity, cosmology*
- *General theoretical physics*
- *Detectors/exp. Techniques*
- *Accelerators, storage ring*
- *Health physics/radiation Effects*
- *Computing/computers*
- *Maths mathematical physics*
- *Astrophysics/ Astronomy*
- *Nonlinear system*
- *Condensed matter*
- *Other fields of physics*
- *Chemical physics/ chemistry*
- *Engineering*
- *Information transfert/Management*
- *Other aspects of science*
- *Economy, commerce, social science*
- *Biography, geography, history*

On peut remarquer que les sujets sont assez divers. Cependant la majorité d'entre eux est en rapport avec la physique des hautes énergies. Il faut donc, en priorité, garder cette ligne de conduite dans la politique d'acquisition. En effet, l'atout du SIS réside dans sa spécialisation, ce serait perdre cet atout que d'entreprendre d'élargir vraiment les thèmes des documents traités.

Le documentaliste n'a parfois pas les qualifications nécessaires pour juger de l'intérêt d'un sujet pour le CERN. C'est pourquoi, au SIS, deux physiciens sont présents pour seconder le documentaliste dans ses choix. Ils travaillent à temps plein dans la service comme documentalistes. Leur double compétence est un atout indéniable pour le SIS.

Voilà donc en quelques mots quelles sont les différentes caractéristiques des importations déjà effectuées. C'est dans cette optique que je devais poursuivre le travail.

¹ <http://library.cern.ch/>

2. Liste des sources explorées

J'ai été amenée à analyser deux types de sources : des bases de données et des compte-rendus de conférences électroniques¹. L'approche est différente pour chaque source ; on les exploite chacune différemment. Par exemple, lorsque l'on analyse une base de données, il faut établir un profil de recherche, car on ne peut pas se permettre d'importer toute la base. En revanche, lorsque l'on veut importer les données de *proceedings* électroniques, on importe généralement tout. La démarche est alors plus simple, il n'y a pas de choix à effectuer à l'intérieur de la base. En effet, lorsque l'on exploite une conférence, il est logique d'importer tous ses *proceedings*.

Mais il est certain qu'un tri sera toujours à faire pour déterminer quelle source est intéressante pour le SIS.

Serveurs de prépublications :

- *Mathematic Preprint Server* : serveur de prépublication créé par Elsevier, il est libre d'accès et permet à tout chercheur de diffuser ses prépublications.
<http://www.mathpreprints.com>
- *Physic Preprint database* : serveur de prépublications en physique de l'UMIST (*the University of Manchester Institute of Science and Technology*)
<http://www.phy.umist.ac.uk/research/preprint.shtml>

Bases de données :

- Les trois derniers mois des enregistrements de la base Pascal, accessibles gratuitement par l'interface ConnectSciences : il s'agit d'une source généraliste dans le domaine des sciences, gérée par l'INIST-CNRS. Elle contient des notices d'articles publiés et de thèses. Pour ne pas confondre cette interface avec la base Pascal complète, nous proposons de la nommer Pascal-ConnectSciences. Le CERN est abonné à la base PASCAL, mais on a choisi d'exploiter cette interface CONNECTSCIENCE, car elle est libre d'accès, il était donc plus facile de l'utiliser. L'intégralité de la base des Articles INIST est accessible par ce portail, cependant l'interface de recherche est peu adaptée. Voilà pourquoi, on s'est contenté d'analyser seulement l'interface qui donne accès aux enregistrements des trois derniers mois de la base Pascal.
<http://connectsciences.inist.fr/>
- *Mpi informatik de Marx-Planck-Institut.fur Informatik*
<http://www.mpi-sb.mpg.de/services/library/catalogues/catalogues.html>
- *Computer Science Bibliography* de l'université TRIER en Allemagne. Il s'agit d'une bibliographie qui concerne les sciences de l'informatique. Elle comporte plusieurs types de documents : les conférences, les revues et les livres. Ce sont les conférences qui nous intéressaient.
<http://www.informatik.uni-trier.de/~ley/db/index.html>

¹ e-proceedings en anglais

Compte-rendus de conférences :

- Liste des *proceedings* de la conférence IEEE Particle Accelerator Conference - PAC '01 , qui s'est tenue du 18 au 22 Juin 2001 à Chicago.
<http://accelconf.web.cern.ch/AccelConf/p01/INDEX.HTM>
- Liste des *proceedings* de la conférence : Particle Accelerator Conference - PAC '99 ; 18th Biennial Particle Accelerator Conference , qui a eu lieu du 29 mars au 2 Avril 1999 à New York.
<http://accelconf.web.cern.ch/AccelConf/p99/PROCS.HTM>
- *Proceedings* de la conférence LINAC 2000 qui s'est déroulée à Monterey du 21 au 25 août 2000.
<http://accelconf.web.cern.ch/AccelConf/l00/index.html>
- *Proceedings* de la conférence de Chamonix en Janvier 2001
<http://sl-div.web.cern.ch/sl-div/publications/chamx2001/contents.html>
- *Proceedings* de la conférence mdk-2001, Genève 26-27 avril 2001.
<http://mdk2001.web.cern.ch/mdk2001/Proceedings/proceedings.html>
- *Proceedings of e+e- physics at intermediate energies workshop*. (30 avril-2 mai 2001 à Stanford (SLAC).
<http://www.slac.stanford.edu/econf/C010430>
- *Proceedings of the 8th International Conference On Accelerator and large Experimental Physics Control System (ICALEPCS 2001)* tenue du 27 au 30 novembre 2001 à San Jose en Californie.
<http://icaleps2001.slac.stanford.edu/>
- *Proceedings of 5th International Symposium On Radiative Corrections (Radcor 2000): Applications of Quantum Field Theory To Phenomenology (11-15 sept. 2000, Carmel, Californie)*
<http://www.slac.stanford.edu/econf/C000911/index.html>

3. La phase d'analyse

« N'est-ce pas là une voie pour l'avenir numérique des bibliothèques ? sélectionner des sources, les analyser, les indexer, les regrouper par thème, pour constituer une collection, correspond bien au savoir-faire des bibliothécaires.[...] Ils apportent une valeur ajoutée bien supérieure à celle des nouveaux outils de recherche, qui indexent d'une manière automatique les sources sans sélection. »¹

Il semble, en effet, que la documentation numérique prenne une place de plus en plus importante dans le schéma de diffusion de l'information. Une des preuves les plus

¹ [5] Le Moal, Jean-Claude / *La documentation numérique - Concurrences et rivalités*, BBF 2002 Paris, t 47 n1. P68

explicités : pendant mes 4 mois de stage au CERN, je n'ai pas eu une seule fois entre les mains un document physique, tous les documents que j'ai eu à traiter étaient numériques. Ceci est emblématique de l'évolution qui est en train de se produire. Voilà pourquoi, au même titre que la tâche traditionnelle d'analyse des documents papiers, l'analyse des sources électroniques devient une activité essentielle dans le travail du documentaliste.

L'analyse que j'ai eu à effectuer n'avait pas pour but d'indexer les sources. Il s'agissait d'un travail d'exploitation, la source est une mine d'informations que je devais sonder. Voilà pourquoi, l'analyse s'est faite en trois temps :

- 1- Analyse informationnelle : l'information des documents est-elle intéressante pour le CERN ?
- 2- Analyse documentaire : les informations bibliographiques sont-elles suffisantes ?
- 3- Analyse matérielle des sources en vue de l'exploitation : l'utilisation de l'*Uploader* nécessite que la source réponde à certains critères.

En effet, il ne s'agissait pas d'intégrer une source dans une liste, mais d'extraire de la source les données intéressantes pour le CERN. Or, il faut que la source analysée respecte certains critères pour que cette opération puisse se faire.

a. Analyse informationnelle

- **Origine de la source**

Quelle université ? institut ? ou organisme ? La provenance de la source indique sa fiabilité.

- **Fréquence de la mise à jour**

Fréquence des nouvelles entrées dans la base, mise à jour de l'interface de recherche... Plus il y a de mises à jour, plus la base est fiable.

- **La qualité des documents**

Quelle est leur valeur scientifique pour le CERN ? Ceci est difficile à déterminer, c'est souvent grâce à la provenance de la source que l'on peut déduire sa valeur. Les physiciens présents dans le service peuvent aider le documentaliste dans cette tâche.

- **Les sujets traités**

Les sujets correspondent-ils aux besoins du CERN ? On doit comparer les sujets proposés dans la source à ceux déjà indexés au CERN. Par exemple, lorsque j'ai analysé la base de prépublications en mathématique, j'ai, avec l'aide d'une physicienne, essayé de trouver les différents sujets adaptés à la demande. Sur 15 sujets, quatre seulement ont été conservés.

- **La langue des documents**

- **Le type des documents.**

S'agit-il d'articles publiés, de prépublications, de thèses... ?

- **La quantité des documents.**

Il est primordial de déterminer la dimension de la source, s'il s'agit d'une base qui contient un grand nombre de documents ou non. En effet, une base qui ne contient que quelques dizaines de documents ne nécessite pas forcément une importation automatique, la saisie manuelle sera plus rapide. Dans le cas contraire, si la base est très importante, il sera nécessaire de faire un choix dans la politique d'importation. Par exemple la base Pascal-ConnectSciences qui est une base généraliste nécessitera la détermination d'un profil pour l'importation. Autre exemple : le *Mathematic Preprint Server* est pour l'instant d'une taille assez réduite, il y a seulement quelques centaines de documents, son exploitation est donc relativement aisée. Plus les sources ont de données, plus elles sont difficiles à exploiter : le tri est plus délicat à faire.

- **Comparaison avec ce qui est déjà présent dans Aleph.**

Si les données de la source explorée sont déjà présentes dans la base, il faut regarder comment elles sont entrées, par quelle source et avec quel statut (Prépublication, article publié...). Par exemple, toujours pour la base Pascal-ConnectSciences, en faisant les tests, je me suis rendue compte qu'on possédait déjà de nombreux articles qu'elle offre, ceux-ci provenant d'INSPEC. Pour cela, il convient de faire des statistiques sur un nombre donné de documents.

Cette première analyse est très importante. C'est en examinant le contenu informationnel de la source que l'on peut effectuer une première sélection.

b. Analyse documentaire

- **Intérêt de la notice :** Quelle est la quantité de l'information fournie ? Tous les auteurs sont-ils cités ? Y a-t-il un abstract, un accès au texte intégral ? On peut comparer une notice incomplète et une notice adéquate :

Exemple de notice de la base de prépublication en mathématiques : *Mathematic Preprint Server*:

Auteur
Titre
Date – Numéro de classement

Les données ne sont pas très nombreuses : il y a seulement l'auteur, le titre et la date. La notice complète avec l'accès au texte intégral se fait par un lien sur le titre (symbolisé par le bleu).

Exemple de notice de la base de prépublication de l'UMIST:

Numéro
Auteur
Titre
Lien vers le texte intégral
Mention de publication
Lien vers une notice plus complète

Cette notice semble plus complète, elle offre le lien direct vers le texte intégral ainsi que la mention de publication. Plus la notice donne d'informations bibliographiques, plus elle est intéressante.

L'accès au texte intégral n'est pas primordial, cependant il reste un critère important dans l'analyse. Par exemple, lors de l'analyse de la *Computer Science Bibliography*, je me suis rendue compte que les accès aux textes intégraux étaient payants, il fallait s'abonner (ACM). Or, l'exploitation de cette source était rendue difficile par un outil de recherche non adapté. J'ai donc décidé de l'abandonner : en effet, il aurait fallu faire un gros travail sur cette source pour ne récupérer que les données bibliographiques, sans texte intégral. Cela aurait été trop peu bénéfique.

- **Qualité de l'outil de recherche.**

Les résultats sont-ils fiables ? Peut-on faire une équation de recherche complexe ? Quels sont les champs sur lesquels on peut faire la requête ? Y a-t-il des opérateurs booléens ? Trouve-t-on des opérateurs de similarité ?

Des sources peuvent être très performantes pour une utilisation normale, la recherche d'un document, par exemple. Mais lorsqu'on souhaite faire un profil de recherche avec une équation assez complexe, cela n'est pas possible, l'interface de recherche n'est pas étudiée pour.

Par exemple, la *Computer Science Bibliography* de l'université Trier offre une interface de recherche qui permet de chercher un document précis, mais il est difficile de faire un profil de recherche plus large. On est donc obligé de travailler chaque conférence présente dans la base une à une, ce qui représente un travail considérable.

En revanche, la base de *preprints* de mathématique (*Mathematic Preprint Server*) est, elle, plus facile à exploiter. En effet, on peut choisir le thème désiré et avoir la liste de tous les documents qui s'y rapportent, le profil de recherche est alors très facile à élaborer. Il suffit de définir les thèmes qui sont intéressants pour les activités du CERN.

- **Y a-t-il la possibilité de créer un profil ?**

Peut-on créer une équation de recherche avec des mots-clés pertinents pour le CERN ? C'est une des phases les plus difficiles et les plus importantes du travail. En effet, si aucun profil ne peut être déterminé, la source sera difficilement exploitable.

On peut créer le profil de plusieurs manières. On peut, dans un premier temps, se servir des sujets indexés dans la source. C'est ce qui a été fait pour le *Mathematic Preprint Server* ; on a déterminé les sujets qui paraissaient importants (*Analysis, Theoretical Computer Science, Computational Mathematic et Probability Theory and Stochastic Processes*).

En revanche pour d'autres sources, il était nécessaire de faire une équation plus précise. On peut se servir, pour cela, des mots-clés utilisés pour la recherche dans INSPEC. Voici, en exemple, l'équation de base qui a été faite pour Pascal-ConnectSciences à partir de ces termes :

(CERN OR LEP OR DELPHI OR CMS OR ALEPH OR ALICE OR CHARM OR CLIC OR COMPASS OR CPLEAR OR FELIX OR ISOLDE OR LARGE HADRON COLLIDER OR LHC OR LHCb OR NOMAD OR OBELIX OR REX OR SPIN MUON OR TILECAL OR TOSCA OR TOTEM OR OPAL OR PROTON SYNCHROTON)

Pour déterminer cette équation il m'a fallu, à partir de la liste des mots-clés utilisés pour INSPEC déterminer ceux qui n'étaient pas pertinents dans Pascal-ConnectSciences. On peut voir le tableau de cette analyse en annexe¹.

- **Pertinence des résultats suite à une requête.**

Il s'agit de vérifier que l'outil de recherche de la source est performant. Dans le cas contraire la démarche n'est pas pertinente. Ce problème s'est posé avec la base Pascal-ConnectSciences. En effet, lorsque l'on inscrivait en requête l'équation ci-dessus, les résultats ne correspondaient pas tous à notre attente. Il fallait d'abord délimiter dans quelle partie de la base était effectuée la requête (quel sujet scientifique était choisi : mathématiques, physique...), et cette fonction n'était pas effective. Par conséquent, la requête était erronée. On peut voir cela dans le résultat d'analyse de Pascal-ConnectSciences en annexe 5.

Une fois que cette première phase d'analyse a été effectuée, et qu'on en a tiré des conclusions positives, on peut se diriger vers un examen plus concret de la source, afin de voir si celle-ci est exploitable.

c. Analyse matérielle : Accès direct aux informations

Le critère de la présentation des informations est très important car l'utilisation de l'*Uploader* ne peut se faire qu'à certaines conditions.

- **Les champs des notices sont-ils correctement structurés ? Y a t-il le respect d'une norme dans le catalogue effectué ?**

En ce qui concerne la structuration des notices, il est nécessaire que celle-ci soit relativement homogène. En effet, s'il n'y a pas une formalisation assez rigoureuse, les données sont difficilement exploitables :

Le champ AUTEUR par exemple peut être structuré de plusieurs façons :

Nom, Prénom
Prénom Nom
Nom, Initiale
Initiale Nom

...

Dans certaines sources, des règles sont mises en place et chaque auteur est catalogué selon la même forme. Mais parfois, aucune règle n'est vraiment respectée, et on trouve plusieurs formes d'écriture. Or, plus il y a de formes d'écriture différentes, plus l'exploitation des données se révèle difficile. En effet, il faut, dans les fichiers de configuration, prévoir tous les cas de figure.

- **Y a t-il un accès direct aux notices ?**

Il faut que les données soient toutes accessibles sur la même page. En effet, lorsque l'on fait une recherche, il est nécessaire que les notices complètes apparaissent sur la page de liste de résultats. Si la page de résultats n'affiche que le titre des documents, cela signifie qu'il y a un lien pour accéder à la notice complète et que toutes les données ne sont pas sur la même page. Ceci n'est pas exploitable. Pour l'*Uploader*, nous utilisons le fichier

¹ Voir annexe 6

texte des résultats (Code source ou enregistrement de la page web en fichier texte). Il est donc nécessaire que les données soient toutes sur cette page (notice complète), car on ne peut pas faire un fichier par notice¹. La récupération ainsi que l'exploitation des données ne seraient pas un gain de temps en comparaison à une saisie manuelle.

Nous avons eu ce problème avec l'interface des résultats d'alerte pour Pascal-Connectsciences : la page de résultats ne donne que les titres, on doit sélectionner un titre pour avoir accès à la notice complète.

- **L'accès au code-source est-il possible ?**

Comme nous l'avons déjà dit, pour récupérer les données des notices, nous utilisons des fichiers textes, car nous ne pouvons nous servir directement de l'interface web. Nous sommes souvent obligés de prendre le code-source HTML. Il faut donc s'assurer que celui-ci est bien accessible, ce qui n'est pas vrai tout le temps. Le code HTML a l'avantage de permettre la séparation des différents champs des notices grâce à ses balises. L'extraction des données est donc plus aisée.

- **Les services proposés sur l'interface.**

On peut, enfin, examiner si l'interface de la source propose des services qui peuvent être utiles pour son exploitation. Il est parfois possible de faire un historique des recherches, cela permet d'éviter de recommencer plusieurs fois les mêmes opérations (gain de temps). De même, il existe des services d'alertes grâce auxquels on peut créer un profil de recherche, ainsi les données nous parviennent de façon régulière au fur et à mesure de leur enregistrement dans la base de données. Cette fonction est très utile, elle permet d'éviter de faire des recherches (annuellement ou plus régulièrement). On constate aussi que ce système permet une plus grande efficacité. La recherche est faite automatiquement, il y a moins de risque d'omettre des informations.

Voici donc les différents critères d'évaluation des sources que j'ai eu à explorer. Il va sans dire qu'aucune n'a jamais rempli tous les critères nécessaires. Il est évident que les différentes bases de données n'ont pas été créées pour pouvoir être utilisées par l'*Uploader*. C'est à nous de nous adapter pour pouvoir les utiliser, puisque l'inverse n'est pas possible. Il s'agit en fait de voir quelles sources remplissent le mieux de critères pour l'exploitation.

Le bilan final du résultat de l'analyse de certaines sources se trouve en annexe².

Sur les 14 sources explorées, seulement 2 n'ont pas abouti à une exploitation.

Il s'agit de la *Computer Science Bibliography* et de la base *MPI Informatik*.

Une fois les sources analysées, il faut alors effectuer un travail plus technique : concevoir les fichiers de configurations du programme *Uploader*.

¹ Voir le schéma en annexe 8

² Voir les annexes 3, 4 et 5.

4. Configuration du programme Uploader

Puisque chaque source est différente, il est nécessaire de créer des configurations propres à celles-ci, pour utiliser l'*Uploader*. Pour cela, on utilise les données originales transcrites en format texte ou visualisées en code HTML. Il est nécessaire de transposer ces données dans un fichier texte, grâce à un éditeur. Comme le programme *Uploader* fonctionne sous UNIX, l'éditeur utilisé est EMACS¹. Suivent ensuite plusieurs étapes :

a. La séparation des données

Le fichier d'origine se compose d'un ensemble de notices, placées les unes à la suite des autres. La première chose à faire est donc de déterminer ce qui permettra au programme de séparer les notices pour les traiter une par une. Il faut donc trouver un élément du fichier qui puisse remplir cette fonction et dont on ne doit pas trouver d'occurrences à l'intérieur même de la notice. Ce choix est capital, il détermine la première étape de l'extraction.

Voici comment se présente le code source des données contenues dans la base de *preprints* de UMIST :

```
<LI>
Preprint: UMIST:Phys:Ast:97-12<BR>
Authors: Isabelle Cherchneff
<BR>
Title: Chemical Models of AGB Winds
<BR>
Published in Molecules in Astrophysics: Probes and Processes, E.F. van Dishoeck
(ed.), 469 <BR>
<A HREF="/cgi-bin/zyview/R=UMIST:Phys:Ast:97-12/D=preprint/V=shv">Detailed
view/Edit</A>
</LI>

<LI>
Preprint: UMIST:Phys:Ast:97-13<BR>
Authors: G. A. Fuller & E. F. Ladd<BR>
Title: Circumstellar Molecular Envelopes <BR>
Published in IAU Symposium 182, Herbig-Haro Flows and the Birth of Low Mass
Stars, B. Reipurth and C. Bertout (eds.), Kluwer <BR>
<A HREF="/cgi-bin/zyview/R=UMIST:Phys:Ast:97-13/D=preprint/V=shv">Detailed
view/Edit</A>
</LI>
```

Nous avons ici deux notices. Elles sont structurées en liste : elles sont donc encadrées par les balises HTML correspondantes . C'est cette dernière balise que j'ai choisie comme séparateur de données. Les notices sont ainsi clairement définies.

¹ Voir en annexe l'exemple d'un fichier de données 9.

Parfois, il n'est pas aussi simple de trouver le séparateur, certaines balises se retrouvant à la fois entre les notices et à l'intérieur de celles-ci. La séparation ne peut alors être efficace. Il faut réussir à trouver l'assemblage adéquat qui ne se retrouve pas au sein même de la notice.

b. Les trois fichiers de configuration

Une fois que la séparation est faite, on peut commencer le travail de transformation des données. Pour cela trois fichiers sont nécessaires¹. Ils doivent être écrits selon une syntaxe particulière à l'*Uploader*. On leur donne des noms significatifs :

*.extract
*.tpl
*aleph.tpl

L'étoile remplace le nom donné à l'ensemble de la configuration. Par exemple, pour la configuration destinée à importer les données de la base de *Preprints* d'UMIST, les configurations se nomment UMIST.extract, UMIST.tpl, UMISTaleph.tpl. Pour des raisons de confidentialité, il n'est pas possible d'expliquer complètement la syntaxe de ces fichiers. Les données de l'exemple qui va suivre sont volontairement simplifiées, les champs sont rarement aussi simples à coder, j'ai seulement voulu montrer la logique adoptée.

➤ Le fichier d'extraction des données : *.extract

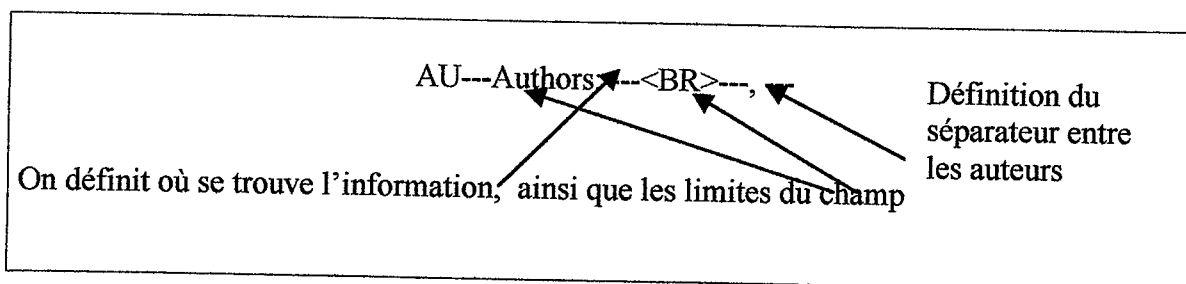
Ce fichier décrit la notice originale pour permettre d'extraire les différents champs qui la composent.

Nous pouvons prendre l'exemple du champ « auteur »

Il se trouve sous cette forme dans la notice originale :

Authors: Isabelle Cherchneff

Dans le fichier *.extract, il est décrit ainsi :



¹ On pourra trouver en annexe 7 le schéma de transformation complet.

A ce stade, l'information que l'on désire importer est extraite et définie.

➤ **Le fichier d'encodage des données : *.tpl**

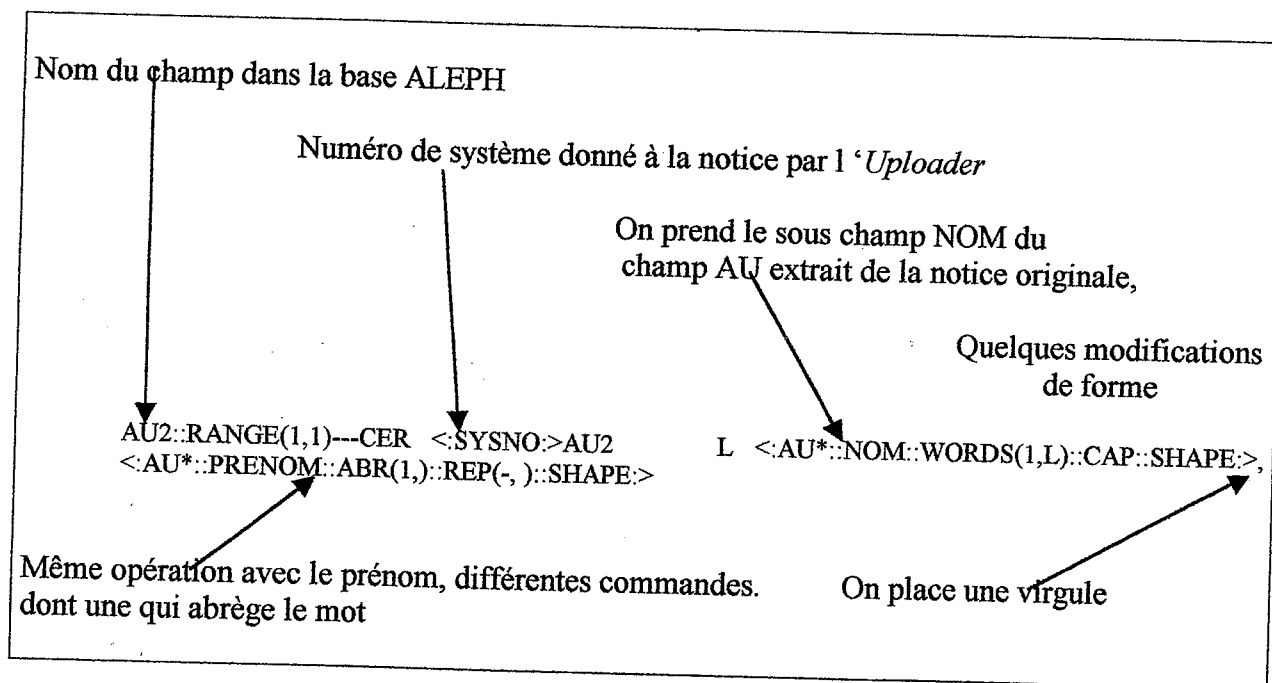
Dans ce fichier, les champs sont encodés pour pouvoir être utilisés ensuite ; on détermine aussi les sous-champs.

AU---<:PRENOM:> <:NOM:>

On définit le fait que le champ auteur est composé d'un nom et d'un prénom.

➤ **Le fichier de structuration de la notice finale : *aleph.tpl**

On crée la notice finale au format utilisé dans la base locale avec les champs extraits grâce aux deux fichiers précédents.



On construit ainsi la notice qui pourra être importée dans ALEPH avec la forme adéquate (Nom, Initiale du Prénom):

Cherchneff, I

Nous n'avons pris qu'un champ dans cet exemple, mais il est évident que les fichiers de configuration décrivent la notice entière.

De nombreuses commandes existent pour créer la notice finale. Il faut maîtriser la syntaxe pour pouvoir les utiliser. Ces configurations peuvent demander un temps de travail important. Il faut prévoir, en effet, tous les cas de figure possibles car il est rare que toutes les notices d'une source soient structurées de la même façon. Dans ce cas, il faut réussir à envisager toutes les possibilités. Le champ auteur est souvent propice à ce genre de difficultés. L'exemple que nous avons montré n'est guère pertinent puisqu'il se révèle être d'une grande simplicité. En effet, il y a souvent plusieurs catalogueurs pour une même source, qui parfois, ne respectent pas les mêmes normes ainsi on peut trouver le champ auteur sous diverses formes :

Nom, Prénom

Prénom, Nom

...

La difficulté s'accroît s'il y a plusieurs auteurs séparés par « and », « & » ou une virgule.

L'*Uploader* permet de déterminer le séparateur pour les auteurs, s'il s'agit d'une virgule, par exemple, on le définit dans le fichier de configuration. Ainsi, le programme peut automatiquement séparer les différents auteurs. Or, j'ai pu constater que très souvent, il y avait plusieurs séparateurs : « and », « & »... dans une même notice. Nous avons donc demandé à Martin Vesely de pouvoir définir plusieurs séparateurs dans la configuration. Il a pu le faire. Ceci est un exemple de la collaboration qui existe entre le support informatique et les documentalistes. Ces derniers déterminent les besoins et les informaticiens créent les applications.

La création de ces fichiers peut donc se révéler être un travail très délicat. Il faut effectuer des tests pour vérifier si les configurations fonctionnent correctement.

Une fois ces fichiers créés, il faut ajouter les caractéristiques de la configuration dans un fichier (main.cfg) pour que l'*Uploader* puisse les utiliser.

On y inscrit le séparateur, le nom de la configuration, les conditions de l'importation...

c. Upload, Correct, Append

Lorsque l'on importe des notices, elles peuvent être utilisées de différentes manières.

- On peut tout d'abord les importer complètement en vue de créer de nouvelles entrées dans la base ALEPH : c'est la fonction *UPLOAD*.
- Il est aussi possible de les utiliser pour mettre à jour des notices déjà existantes dans ALEPH. Par exemple, pour la mise à jour de notices de *preprints* par l'ajout des

références de publication, on corrige le nom de la base (les *preprints* sont en base 11, les articles publiés sont en base 13), on corrige alors la date et les mentions de publication. C'est la fonction *CORRECT* ; il faut alors créer un fichier de configuration spécial qui définit les différents champs à corriger. (*correct.grp)

- Enfin, on peut, grâce aux nouvelles informations apportées, ajouter des champs dans des notices déjà existantes. C'est la fonction *APPEND*. On peut, par exemple, ajouter un lien vers le texte intégral. (Création du fichier *append.grp)

On peut utiliser ces trois fonctions à la fois. L'*Uploader* cherche dans ALEPH si la notice y est déjà présente grâce à une équation de recherche que l'on définit dans le fichier main.cfg. Il faut, dans cette équation, faire en sorte que la recherche soit la plus efficace possible.

Ainsi, pour la configuration UMIST, j'ai choisi comme équation : chercher le nom de l'auteur et les quatre premiers mots du titre. Cette équation doit être, elle aussi, écrite selon une syntaxe spéciale. Pour cette recherche, il faut utiliser les éléments les plus pertinents. En ce qui concerne, par exemple, les comptes-rendus de conférence RADCOR00 ; comme on possédait déjà dans la base les *preprints* soumis à Los Alamos, un numéro de serveur unique présent à la fois dans la notice d'origine et dans ALEPH a servi d'élément de recherche. La fiabilité en était plus grande.

d. Les "knowledge bases"

Lorsque l'on importe des notices bibliographiques dans la base ALEPH, il est important que les informations contenues soient normalisées. A ce jour, aucune norme n'est vraiment appliquée au niveau international, chaque organisme ayant ses propres codes. Il est donc nécessaire de normaliser les données avant de les intégrer dans la base ALEPH. C'est dans ce but que des listes de référence ont été créées.

Par exemple, il existe un fichier qui permet de transcrire tous les titres de périodiques dans une forme abrégée normalisée (utilisation de la norme ISO 4). Cela évite que le même périodique apparaisse sous plusieurs formes. L'indexation et la recherche en sont grandement facilitées. Cette *knowledge base* permet, par ailleurs, d'activer le lien vers la publication électronique de la revue (si le CERN est abonné).

Par exemple, le titre du périodique Physical Review peut se trouver sous différentes formes selon les sources : *Phys. Review*, *Physical Review*, *Phys Rev*... dans le fichier *knowledge base*, toutes ces formes sont énumérées pour pouvoir être transformées dans la forme normalisée : *Phys. Rev*.

Pour faire fonctionner ces listes de références, il suffit de les appeler à partir du fichier de configuration *aleph.tpl.

Parfois, il est nécessaire de créer une *knowledge base* particulière, propre à la configuration : pour transformer la structure des dates, par exemple. Pour la configuration UMIST, j'ai été amenée à effectuer des *Knowledge Bases* pour le champ Année et le champ « Sujet ».

Exemple de KB pour normaliser la forme des années :
01 → 2001 02 → 2002 ...

Ces *knowledge bases* évitent de faire des modifications trop longues à l'intérieur des fichiers de configuration. Par ailleurs, elles peuvent desservir d'autres sources.

e. Le texte intégral

Ce programme, particulièrement intéressant, permet une importation massive d'informations dans un temps relativement court.

Il est temps de s'arrêter sur la nature des données bibliographiques importées et notamment sur le texte intégral.

Souvent les notices sont accompagnées d'un lien vers le texte intégral. Il est intéressant pour le SIS d'importer ce lien, il permet, en effet, au chercheur d'accéder directement au document. Cette importation est tout à fait légitime puisqu'on n'extrait que le lien (l'URL) et non le texte lui-même. J'ai pu effectuer cette opération pour de nombreuses sources. L'inconvénient de ce procédé réside dans le fait que les URL sont très souvent instables, et peuvent être changées. Alors, le lien n'est plus valide, cela pose un problème de fiabilité.

Or, il existe une fonction sur le programme qui permet de remédier à cette lacune. L'*Uploader* permet, en effet, d'importer directement, lorsqu'il y a lieu, le texte intégral des notices sur le serveur du CERN.

Il s'agit de la fonction DOWNLOAD que j'ai eu l'occasion de manier lors de mon travail sur les *proceedings* des conférences PAC 1999 et PAC 2001. Cette fonction n'étant pas encore finalisée, il a fallu faire appel au support informatique. Après quelques essais non concluants, nous avons réussi à activer cette commande.

Ainsi, les documents ne sont plus seulement accessibles par un lien vers l'emplacement dans la base de données d'origine, ils se trouvent directement stockés sur le serveur du CERN. Cette manipulation évite les problèmes inhérents à l'instabilité des adresses URL. Bien entendu, cette importation du texte intégral ne peut se faire sans l'autorisation préalable de l'organisme qui le possède. Mais, les documents importés étant en général des *preprints*, il n'y a pas de problème de droit d'auteur.

En ce qui concerne les formats des textes intégraux, on remarque que le format lisible sous Acrobat Reader : Portable Document Format est le plus répandu. Il remplace de plus le format Post Script. On trouve parfois des documents Word (.doc ou .rtf), mais c'est assez rare, les scientifiques utilisent d'avantage le format TeX que Word, il est, en effet, beaucoup plus adapté à l'écriture des formules mathématiques.

5. Importation et validation des informations importées

Une fois que le programme est configuré et que tous les tests se révèlent être concluants, on peut passer à la phase d'importation.

Il faut tout d'abord récupérer les données dans la source grâce à l'équation de recherche définie lors de la phase d'analyse. Parfois quelques modifications sont à faire dans le fichier de données (retirer les éléments superflus...). Puis, on fait tourner le programme *Uploader*.

Lorsque les notices ont été transformées, un autre type de travail s'impose : la correction. Le programme met à part les notices qui lui semblent incorrectes (*Non confident*), mais il est tout de même nécessaire de contrôler toutes les notices transformées. Pour l'instant, il est encore impensable d'intégrer les données transformées par l'*Uploader*

dans ALEPH sans vérification préalable. Il est souvent nécessaire de vérifier l'un des champs.

Pour l'importation des PAC 1999, les champs Auteurs n'étaient pas fiables. Cela était dû au manque de rigueur dans l'inscription des noms d'auteurs dans les notices. La configuration ne pouvait pallier à cette lacune. Il fallait donc vérifier les champs AU et corriger ceux qui étaient défectueux. Pour cela, il existe une méthode : il suffit de créer un champ AUBIS qui importe le champ Auteur de la notice initiale sans le formater, ainsi on peut comparer les deux champs, et voir rapidement ce qui a été mal transcrit.

Ce travail est répétitif et laborieux ; malheureusement il reste nécessaire, car la quantité des informations ne doit pas primer sur leur qualité. Une information erronée peut se révéler très pénalisante dans l'utilisation ultérieure de la notice (Indexation, recherche...).

Cependant cette phase de correction ne doit pas occulter le reste des possibilités qu'offre le programme *Uploader*. Il permet, en effet, d'améliorer considérablement le travail d'importation de données. Ainsi, le passage de l'informel au formel demande un temps moins important, c'est ce qu'on doit retenir. L'élimination de toute erreur est dans l'absolu impossible, on doit seulement essayer d'en réduire au maximum le nombre. On peut, par exemple, évoquer le programme de transformation de données créé par Mathilde de Saint Léger. Ce programme prévoyait un fichier de notices incorrectes. La transformation n'était pas complètement automatisée.

6. Bilan des importations¹

Sources	Type de documents	Etat de l'importation	Nombre de notices importées	Date
BASES DE DONNEES				
UMIST (Manchester)	<i>Preprints</i>	De 1995 → 2001	72 (nouveaux)	Mai 2002
MATHPREP serveur Elsevier	<i>Preprints</i>	Annuelle	61 (nouveaux)	Juin 2002
PASCAL ConnectScience	Articles	A tester ultérieurement	102 (mises à jour publications)	Mai 2002
PROCEEDINGS				
PAC 99 (<i>e-proceedings</i>)	Articles	Terminé	1194 (235 mises à jour publications + 959 nouveaux)	Mai 2002
PAC 2001 (<i>e-proceedings</i>)	Articles	Terminé	1304 (188 mises à jour publications + 1116 nouveaux)	Mai 2002
LINAC 2000 (<i>e-proceedings</i>)	Articles	Terminé	323 (226 mises à jour publications + 97 nouveaux)	Juillet 2002
EEPIE 2001 (<i>e-proceedings</i>)	Articles	Terminé	37 (18 nouveaux, 19 mises à jours)	Juillet 2002
ICALEPCS01 (<i>e-proceedings</i>)	Articles	Terminé	205 (63 nouveaux, 142 mises à jours)	Juillet 2002
CHAMONIX XI (<i>e-proceedings</i>)	Articles	Terminé	71	Juillet 2002
MDK 2001 (<i>e-proceedings</i>)	Articles	Terminé	33	Juillet 2002
RADCOR 2000 (<i>e-proceedings</i>)	Articles	Terminé	55 mises à jour	Juillet 2002

Voici l'ensemble des résultats auxquels mon travail a abouti :

- Exploitation de 11 sources
- 3457 notices importées.

A cela s'ajoute l'importation de 2654 notices d'une sources déjà configurée : PROQUEST²

En ce qui concerne la nature des données importées et notamment le texte intégral, nous pouvons consulter le tableau récapitulatif suivant.

¹ Voir le détail de chacune dans la partie II 2. Les sources explorées.

² Base de données de thèses.

Sources	Texte Intégral
UMIST	Lien vers le texte intégral
MATHPREP	Lien vers le texte intégral
PASCAL-CONNECTSCIENCES	Pas de texte intégral
PAC 99	Importation du texte intégral
PAC 01	Importation du texte intégral
LINAC 2000	Lien vers le texte intégral
EEPIE 2001	Lien vers le texte intégral
ICALEPCS01	Lien vers le texte intégral
CHAMONIX XI	Importation du texte intégral
MDK 2001	Importation du texte intégral
RADCOR 2000	Pas de texte intégral

Le bilan est assez positif puisque seulement deux sources n'ont pas permis l'accès au texte intégral. (RADCOR 2000 et Pascal-ConnectSciences).

Ainsi s'achève cette partie de description de la mission. Toutes les étapes ont été expliquées. Il nous faut désormais essayer de voir quelles conclusions on peut tirer de ce travail.

III. Bilan et perspective pour cette forme de traitement automatique du document.

Il nous faut maintenant, au risque de paraître un peu académique, tenter de faire un bilan de ce travail effectué pendant les quatre mois de stage. Il est temps, au vu des résultats obtenus, de prendre du recul vis-à-vis de cette forme de traitement du document et d'en faire un état des lieux.

Tout d'abord, nous pourrions déterminer, à la lumière des problèmes rencontrés lors du travail, quels sont, au niveau de la simple utilisation, les atouts et les limites d'un tel programme.

Nous pourrions voir aussi quelles sont ses influences sur la politique d'acquisition de la littérature grise au CERN. L'*Uploader* est, en effet, utilisé depuis deux ans déjà, il est temps de le mettre en perspective et de faire un bilan des importations de littérature grise effectuées au CERN depuis deux ans.

Après une telle analyse, il conviendra de voir quel est l'avenir d'un tel programme dans un contexte où germe sans arrêt une nouvelle initiative sur la normalisation de l'information.

1. Intérêts et limites du programme Uploader

Faisons un point tout d'abord sur les aspects positifs et négatifs que j'ai pu déceler au cours du travail effectué grâce à l'*Uploader*. Cet outil a été la source d'une nette amélioration du nombre des importations de documents dans la base du SIS ; cependant, il est nécessaire de noter que ce programme a aussi quelques inconvénients.

a. Intérêts

- L'augmentation du nombre des importations

L'utilisation de l'outil *Uploader* a été un vrai progrès pour le fonds de littérature grise au CERN. Le nombre des importations a considérablement augmenté. Il est évident que comparativement à la saisie manuelle de données, l'*Uploader* est beaucoup plus rapide et efficace. En effet, une importation automatique peut engendrer un nombre très important d'enregistrements en une seule fois. Par exemple, lorsque j'ai importé les données de la base PROQUEST, plusieurs milliers de notices ont été créées. On peut analyser les statistiques des importations depuis quelques années pour constater l'essor du nombre d'enregistrements depuis le début de l'utilisation de l'*Uploader*¹.

- le déplacement des tâches du documentaliste

On constate plus précisément un déplacement des tâches. En effet, la saisie manuelle est à la portée de tout le monde ; c'est un travail qui ne demande pas de qualification précise, si ce n'est de la rigueur et de la précision. En revanche, l'utilisation de l'*Uploader* nécessite une connaissance approfondie de cet outil. La confection des configurations demande un apprentissage des commandes à utiliser ainsi que de la syntaxe. En comparaison avec la saisie manuelle, l'intelligence et la logique du documentaliste sont mises à l'épreuve. Même si l'utilisation de l'*Uploader* réduit considérablement la contrainte du facteur 'temps', elle ne l'élimine pas totalement: parfois, réussir à faire une

¹ Voir l'annexe 10, les statistiques de Catherine Cart

configuration parfaite demande plusieurs jours de tests et de tâtonnements. Seulement, il rend le travail beaucoup plus intéressant. On peut dire aussi que cet outil implique un changement du travail du documentaliste, il n'est plus seulement l'intermédiaire physique entre le document et l'utilisateur, mais prend en charge, en quelque sorte, une partie du travail de l'utilisateur.

Par ailleurs, les configurations deviennent un acquis, car elles peuvent servir pour d'autres sources. La saisie manuelle est un acte premier qui n'engendre rien d'autre que la saisie même. Il doit sans cesse se répéter pour que la saisie continue. Au contraire, le travail de configuration de l'*Uploader* pour une source donnée devient un outil dont on peut se resservir, il devient un acquis. Si, à cause de la confection de ses configurations, l'*Uploader* n'est parfois pas plus rapide que la saisie manuelle, c'est dans la réutilisation de ces acquis qu'il permet de créer un gain de temps indéniable. Par exemple, après avoir effectué les configurations pour la conférence PAC 1999, j'ai pu les réutiliser en les modifiant un peu pour PAC 2001.

- Une plus grande fiabilité

Un autre avantage indéniable de ce programme réside dans le fait que comme le procédé est automatique ; il y a peu de chance que la machine commette des erreurs en dupliquant les noms d'auteurs ou les titres. On peut donc dire qu'il y a une plus grande fiabilité dans la conservation des données. En effet, la saisie manuelle peut être source d'erreurs, de fautes de frappes que la machine, elle, ne fait pas.

- Création d'une valeur ajoutée

Lors de la transformation des données d'une notice, on peut ajouter certains champs, et donc rendre la notice plus complète. Par exemple, pour les contributions de conférence, on ajoute le champ LKR, qui donne les informations de publication du compte-rendu et permet de lier toutes les contributions à une même conférence.

- Une autre conception de la politique d'acquisition de la littérature grise

L'*Uploader* a permis, en outre, de mettre en place une réflexion générale sur la politique d'acquisition de la littérature grise au CERN. Avant, il n'y avait pas de choix possible, la bibliothèque était réduite à n'enregistrer dans sa base que ce qu'elle recevait par les *mailing-lists*, elle était dépendante de ce vecteur de diffusion de l'information. Or ces *mailing-lists* n'étaient pas toujours complètes : Le SIS n'avait pas accès à toute l'information. Aujourd'hui, c'est la démarche inverse qui s'opère : c'est le SIS qui va chercher l'information où elle se trouve, il n'est plus obligé de l'attendre. Alors qu'auparavant, on se contentait de ce qu'on nous donnait, désormais, on peut tenter d'acquérir toute l'information désirée. Cela change beaucoup de choses dans les perspectives d'acquisition.

La politique principale est d'archiver tous les papiers qui proviennent du CERN ou qui en parlent, mais avec l'*Uploader*, le champ d'action est considérablement agrandi. Cela implique une nécessaire réflexion. On ne peut pas non plus se permettre de faire de l'importation massive sans garder quelques critères de sélection. Nous vivons, en effet, à une époque où l'information est partout, les bases de données (gratuites ou payantes) fleurissent. Il est donc nécessaire de faire un tri de cette information, c'est ce que j'ai été amenée à faire au cours de mon travail. Toutes les sources qui se présentent à nous ne peuvent être exploitées, d'une part pour une question de temps et d'autre part, parce qu'il faut se souvenir que le public auquel est destiné la base est un public spécialiste. Le CERN est un laboratoire de recherche en physique des particules et nous ne pouvons nous permettre d'agrandir le champ des sujets traités. Ce serait perdre la spécificité de la base de

données. En effet, il faut savoir garder une certaine ligne de conduite dans la politique d'acquisition, on risquerait, dans le cas contraire, de se perdre dans «le trop d'information». C'est ici que les choix deviennent délicats à effectuer, comment prévoir ce qui est susceptible d'intéresser un physicien du CERN ? Certains critères sont à prendre en compte, bien entendu, certains éléments sont reconnus comme critère pertinent pour le CERN (Université d'origine, source, affiliation des auteurs...). Mais il est vrai que dans un contexte où le *self-archiving* se développe très rapidement, les critères d'évaluation classiques (Comité de rédaction, *Peer-review*) sont de moins en moins pertinents puisque ce n'est plus grâce à eux que l'on juge si un document doit être diffusé ou non. Ce qui signifie que l'on peut trouver, sur les serveurs de prépublication, tout aussi bien des documents intéressants que des documents qui n'ont pas de réelle valeur scientifique. Les critères d'acquisition doivent donc être définis autrement.

L'*Uploader* rend donc plus intéressant le travail d'importation, à tous les niveaux que ce soit. Il est évident qu'il apporte une valeur ajoutée aux tâches de la saisie de données dans la base HEP, et cela, parce qu'il ouvre, à la politique d'acquisition, de nouvelles perspectives. Auparavant, avec les *mailing lists*, elle était passive, désormais, elle est dynamique.

b. Les limites

Cependant, nous ne pouvons décrire cet outil sans parler aussi de ses limites.

- *L'Uploader reste un programme d'automatisation partielle*

Rappelons tout d'abord qu'il s'agit d'un outil de traitement semi-automatique du document, la procédure n'étant pas complètement automatisée. L'intervention de l'homme est, en effet, nécessaire au début et à la fin de la chaîne de traitement. Tout d'abord, pour chaque source, il faut faire une configuration différente, il faut donc à chaque fois adapter l'outil à la source à exploiter. Cette démarche peut être un peu longue parfois. A la fin de l'importation, une intervention est souvent nécessaire pour contrôler les données résultantes. Parfois certains champs ne peuvent pas être correctement transformés : cela se voit surtout pour les sources dans lesquelles le catalogage est peu normalisé et où les champs ne sont pas tous structurés de la même façon. En effet, il arrive que les commandes de l'*Uploader* utilisées dans les configurations ne permettent pas de faire une notice parfaite. Il est alors nécessaire de corriger les champs qui n'ont pas pu être formatés correctement.

- *Certaines rigidités du programme*

Par ailleurs, on constate aussi que l'outil reste assez rigide. Il ne peut pas s'adapter à toutes les situations et donc reste fortement dépendant de la source que l'on veut exploiter. Or lorsque celles-ci présentent certains défauts, l'*Uploader* reste impuissant.

- *Limitation des entrées dans ALEPH*

Enfin, il faut parler du fait que si l'*Uploader* permet d'importer un nombre très important de notices, les importations dans ALEPH sont, elles, limitées. On ne peut dépasser un certain nombre de notices importées par soir, les informations étant chargées la nuit dans ALEPH. Il résulte donc de la transformation massive de données effectuée par l'*Uploader*, un nombre important de notices qui parfois est arrêté par le goulet du passage dans ALEPH. On se retrouve alors avec des données en attente qui ne peuvent être transférées directement dans la base. Il se peut qu'avec le passage au nouveau système, cet

inconvenient soit atténué. Pour l'instant, il est nécessaire de réguler le flux des entrées de données. Ceci peut parfois poser certains problèmes, cela ralentit d'une certaine manière les importations. Pour l'importation des données de PROQUEST, le problème s'est posé. Cette base contient un nombre très important de données. J'ai extrait plus de 5000 notices en avril, au début de mon stage. Quand je suis partie, en juillet, seulement 3000 étaient versées dans ALEPH, 2000 d'entre elles étaient toujours en attente.

L'*Uploader* est donc un outil très performant mais qui garde certaines contraintes. Cependant, les inconvenients ne doivent pas non plus cacher ses intérêts. Il nous faut maintenant faire un bilan plus précis des importations effectuées pendant le stage. Certaines remarques sont à noter.

2. Un regard sur les sources exploitées.

Le bilan que l'on peut faire des importations¹ effectuées ne peut se faire sans la prise en compte de l'état des sources que j'ai eu à exploiter. Voici le tableau récapitulatif :

Sources exploitées	Sources abandonnées après analyse	Sources abandonnées après test
Mathematic preprint server	Mpi informatik	Pascal-Connectscience
Physic Preprint database	ComputerScience Bibliography	
PAC01 (<i>Proceedings</i>)		
PAC99 (<i>Proceedings</i>)		
LINAC 2000 (<i>Proceedings</i>)		
Chamonix XI (<i>Proceedings</i>)		
MDK 2001 (<i>Proceedings</i>)		
e+e- physics at intermediate energies workshop (<i>Proceedings</i>)		
ICALEPCS 2001 (<i>Proceedings</i>)		
RADCOR 2000 (<i>Proceedings</i>)		

Sur quatorze sources, seulement trois n'ont pas abouti à une importation. On peut dire que le bilan est plutôt positif. On compte 3457 notices importées (nouvelles créations et mises à jour). Seulement, si on analyse un peu, on se rend compte que la plupart des notices proviennent principalement de deux sources PAC 99 et PAC 01. Le reste des sources n'apporte que très peu d'enregistrements par rapport à ce qui était espéré. L'exemple de Pascal est le plus pertinent : on compte l'importation de 102 notices seulement, alors que cette base est très importante. Cette source n'est pas la seule à fournir ce résultat.

On peut constater que les petits serveurs et les bases de *proceedings* ont été plus

¹ Voir le tableau des importations situé dans la partie II

adaptés que les sources beaucoup plus importantes à l'utilisation de l'Uploader.

Le bilan des importations est donc convenable mais il ne correspond pas exactement à nos attentes. Nous pouvons tenter de fournir quelques hypothèses pour expliquer cela.

a. Il existe des sources non adaptées à l'Uploader.

Dans ce processus d'importation, toutes les sources ne sont pas « bonnes ». En effet, les bases de données bibliographiques gratuites ou payantes sont de plus en plus nombreuses sur le web. Chacune a ses objectifs propres, et il est normal que chacune soit conçue en fonction de ses objectifs. On constate, par exemple, que la base Pascal accessible par le portail ConnectSciences ne présente pas tous les critères nécessaires pour pouvoir être traitée grâce à l'Uploader :

- Pas d'accès au texte intégral, il faut payer.
- Interface de recherche non adaptée pour faire un profil de requête exploitable pour l'importation massive.
- L'accès aux notices données par les alertes se fait sous forme de liens, il n'y a qu'une notice par page web.

...
Tout cela empêche une exploitation correcte de la source. Cependant, les concepteurs de la base Pascal ne l'ont pas créée pour qu'elle puisse être utilisée par le CERN. Cette base réalisée en partenariat avec différents organismes nationaux et internationaux a un but commercial : elle est accessible sur serveur ou sur CD ROM. Elle est en accès libre seulement sur ConnectSciences. Ce portail n'a donc pas pour but de diffuser gratuitement l'information, il donne un échantillon de la base de données entière, peut être pour conquérir de nouveaux clients. La base Pascal-ConnectSciences est, en effet, payante : l'outil de recherche permet de déterminer quel(s) article(s) on cherche, et si celui-ci nous semble intéressant, on le commande à l'INIST. Cet aspect commercial rend la source inexploitable au CERN, on ne peut pas dire pour autant qu'elle soit déficiente.

b. Un manque de stabilité de certaines sources.

J'ai noté que certaines bases de données assez importantes sont difficilement exploitables à cause de leur instabilité.

Nous pouvons prendre l'exemple de Compendex¹. Cette base, gérée par Elsevier, est une des plus importantes bases de données mondiales de documents scientifiques, techniques et médicaux offrant le texte intégral. Elle contient 40 millions de résumés d'articles scientifiques et possède des liens vers plus de 80 sites d'autres éditeurs de journaux scientifiques. Or, les tests d'importations ont échoué à cause de l'instabilité de la base. Les champs des notices souffrent d'un manque de normalisation évident, ce qui a rendu la création d'une configuration de codage de données relativement délicate. Par ailleurs, la recherche manque de fiabilité et l'indexation est quelque peu déficiente, ce qui rend le résultat des requêtes non pertinent.

J'ai constaté aussi des problèmes similaires lorsque j'ai analysé la base Pascal-ConnectSciences (outil de recherche peu fiable, services proposés non fonctionnels...), cela peut s'expliquer par le fait que celle-ci a été créée récemment. D'ailleurs, j'ai pu faire

¹ Source analysée par une autre stagiaire qui a pu remarqué ces déficiences.

différentes analyses au cours de mon stage, et une évolution notable était visible¹. On peut donc penser que, dans un futur proche, cette source sera plus stable.

On remarque donc que certaines grandes bases qui jouissent d'une certaine renommée n'offrent pas les prestations attendues, alors que parfois les petites bases plus intimes sont bien plus fiables. Mais ce constat ne doit pas devenir une généralité.

c. Une redondance des données

On peut noter, par ailleurs, que les informations contenues dans différentes sources ont tendance à se recouper. Ainsi, la plupart des notices des articles contenus dans Pascal-ConnectSciences sont déjà importées par l'intermédiaire de la base de données INSPEC au CERN. On remarque que, sur toutes les notices issues de Pascal-ConnectSciences, mais qu'on ne possède pas dans la base du CERN, la plupart sont contenues dans INSPEC ; on les recevra donc par cet intermédiaire. Ce recoupement de données est normal puisque ces sources apportent le même type de documents, elles puisent, en effet, leurs données des mêmes périodiques.

Ceci est à prendre en compte dans le travail d'importation. Ce constat permet, en effet, d'éviter de fournir des efforts inutiles.

d. Quel bilan peut-on conclure de tout cela ?

Face à tous ces constats énumérés ci-dessus, on peut se demander quelle est la conclusion que l'on peut en tirer. Avant tout, regardons le tableau de statistique de Catherine Cart qui montre l'évolution des importations automatiques entre 2000 et 2001. On constate une légère diminution du nombre des enregistrements. On pourrait en déduire que l'on arrive à un état d'essoufflement des sources. C'est vers cette hypothèse que pourraient mener les différentes remarques notées plus haut.

En effet, on exploite beaucoup de sources très importantes qui couvrent déjà une grande partie de l'information scientifique. Les autres sources moins importantes sont-elles superflues ? Parfois, ces dernières n'apportent que quelques documents². On se demande alors si une importation automatique se justifie. Pour quelques dizaines de documents, un catalogage manuel semble plus approprié.

Il est vrai que certaines bases explorées n'ont pas apporté le résultat espéré, du fait d'une redondance des données avec d'autres sources. C'est le cas pour la base Pascal-ConnectSciences.

Cependant, cet état des lieux est difficile à faire à l'heure actuelle. Nous n'avons pas, en effet, assez de recul pour faire un vrai bilan. De plus, si le nombre des importations a été supérieur en 2000, cela s'explique par le fait que 2000 a été la première année d'utilisation de l'*Uploader*. On a donc fait des importations rétrospectives de données enregistrées avant cette année-là. La comparaison avec 2001 est donc faussée. On ne peut tirer de réelles conclusions à partir de ces chiffres.

¹ Voir en annexe 5 les deux résultats d'analyse de Pascal-ConnectSciences

² Voir le tableau des importations dans la partie II.

Il faudra attendre encore un ou deux ans pour déterminer si on arrive à un essoufflement des sources, si nous avons atteint une exploitation optimale de l'ensemble des sources documentaires.

3. Uploader : combler un manque en attendant mieux ?

A l'origine, l'*Uploader* a été conçu pour combler un manque. Il s'agissait de récupérer les documents issus du CERN que les chercheurs oublièrent de soumettre au SIS. L'objectif premier de cet outil a vite été dépassé. L'*Uploader* a, en effet, permis de récupérer un nombre considérable de données d'autres instituts. Il permet de transformer les notices de ces sources dans le format adéquat pour ALEPH. Tous les systèmes documentaires ont une notice minimale constante, mais on constate cependant un manque de normalisation des différents catalogages. C'est ce manque que l'*Uploader* visait à combler. Or, on peut voir que les sources d'auto-archivage telles que le serveur XXX de l'université de Cornell sont de plus en plus importantes. Ce sont désormais les chercheurs qui soumettent eux-même leur *preprints*. Et c'est ici que le manque de normalisation est le plus visible. Les chercheurs ne sont pas des documentalistes...

Voilà pourquoi depuis quelques années de nombreuses initiatives ont vu le jour pour compenser cette carence.

On peut parler, par exemple de l'*Open Archive Initiative*¹. Ce projet est la réponse à un appel en juillet 1999 de Paul Ginsparg (initiateur de la base de *preprints e-Print archive* à Los Alamos), de Rick Luce et Herbert Van de Sompel (Los Alamos). Leur souhait était de créer un service universel d'auto-archivage des publications scientifiques par leurs auteurs. Désormais, l'objectif de l'*Open Archive Initiative* est de favoriser la diffusion de l'information scientifique et technique (en particulier les *preprints*). Pour cela, différentes bibliothèques et centres de documentation tentent d'adopter des normes communes afin d'échanger facilement des données sans de lourdes modifications.

Par exemple, le *Mathematics Preprint Server* (Elsevier) devrait dans un futur assez proche être reconnu par l'*Open Archive Initiative* comme un fournisseur de données contrôlées.

Ces initiatives avancent, pas à pas, mais il est vrai que l'ampleur des travaux à entreprendre est considérable. Le CERN, lui-même, participe activement à cette initiative. L'idée commence à faire son chemin.

Or, si toutes ces démarches aboutissent, l'*Uploader* perdrait sa fonction principale. Il n'y aurait, en effet, plus besoin de transformer les données importées. L'*Uploader*, dans le contexte actuel, n'est pas destiné à avoir une durée de vie éternelle. Il n'est là que pour combler provisoirement un manque amené à disparaître.

On peut dire toutefois, que ces initiatives de normalisation de l'information n'en sont qu'à leur début. L'idée semble pour l'instant utopique. L'*Uploader* a donc encore du temps devant lui... Il reste, pour l'instant un outil indispensable.

¹ Pour en savoir plus, on peut consulter le site officiel de l'OAI : <http://www.openarchives.org> ou la page du site du CERN consacrée à ce sujet : <http://documents.cern.ch/OAI>

4. Une remise en cause du système traditionnel

Avant de conclure, il me semble important de faire une remarque. Il faut, en effet noter que ce système de traitement des *preprints* que nous venons d'étudier, et plus généralement le développement des serveurs de prépublication entraînent une évolution du panorama traditionnel de la communication et de l'édition scientifique. On a pu constater que les *preprints* prenaient une place de plus en plus importante. Cela implique une transformation significative du schéma de diffusion de l'information scientifique.

En effet, on constate que le système habituel : auteur – éditeur – bibliothèque-lecteur n'est plus le seul valable. Puisque les chercheurs peuvent désormais diffuser eux-même, par l'intermédiaire des serveurs, leurs propres résultats de recherche, le poids de l'éditeur se fait plus faible. On obtient le schéma suivant :

Auteur – serveur – lecteur.

Il est vrai que l'éditeur retrouve son rôle lorsque l'article est publié, et il semble que la publication soit amenée à subsister puisqu'elle implique le passage de l'article devant un comité de lecture, et donc une validation de cet article. On constate cependant que les serveurs de *preprints* laissent l'article accessible même après sa publication. Ainsi, le *preprint* n'est plus seulement un état provisoire du texte. « *Les prétirages électroniques représentent une source d'information de plus en plus importante pour les chercheurs. Les serveurs de preprints sont souvent le premier choix des physiciens, astrophysiciens ou encore mathématiciens pour trouver des informations sur des recherches en cours ou suivre les travaux de leurs collègues.* »¹ Par ailleurs, on constate que, dans certaines revues très renommées, des *preprints* sont cités en référence. Ce type de document devient donc un vecteur d'information à part entière.

Ce phénomène est source de querelles importantes entre les éditeurs et les serveurs de prépublication. « *Les éditeurs voudraient obliger les auteurs à retirer leurs articles de ces bases lorsqu'ils sont effectivement publiés dans leurs revues, mais le rapport de force est tel que certains éditeurs ne s'y risquent plus.* »²

L'importance de cette évolution peut être illustrée par le fait que même des éditeurs s'appliquent à favoriser la création de serveurs de *preprints*. J'ai étudié, en effet, le *mathematic preprint serveur*. Or, celui-ci est une initiative d'Elsevier, un des plus grands éditeurs scientifiques du monde. Il reste à se demander quelle est la rentabilité financière qu'ils en tirent.

Il est important de noter que les chercheurs saisissent eux-même leurs travaux sur les serveurs de prépublication, il y a de moins en moins d'intermédiaire dans le système de diffusion. Les auteurs sont responsables de l'enregistrement de leur données bibliographiques. Or, comme nous l'avons vu plus haut, ce ne sont pas des documentalistes, c'est pourquoi, certains serveurs de *preprints* ont des systèmes de saisie qui permettent de normaliser les données. Il sera donc toujours nécessaire d'avoir des

¹ [7] Pignard, Nathalie ; Paillart, Isabelle (dir.) / *Les nouvelles formes de publication scientifique sur Internet - La remise en cause du modèle éditorial traditionnel*. Juin 2000. p 74.

² [2] Chartron, Guislaine / *La presse scientifique sur les réseaux*. 1995, Revue électronique Solaris : <http://www.Unicaen.fr/bnum/jelec/Solaris>

programmes pour transformer les données, et ainsi passer de l'informel au formel. Des programmes comme l'*Uploader* seront donc toujours utiles, mais peut-être, seront-ils exploités différemment.

Voici donc le bilan que j'ai pu tirer de mon travail sur cette forme de traitement automatique de documents. Elle a ses points forts et ses faiblesses, mais elle reste un atout indéniable du Service d'Information Scientifique du CERN. Une conclusion définitive sur l'avenir de l'*Uploader* ne peut encore être faite. Il faudra attendre quelques années pour déterminer quelle est la pérennité de ce système. Il est, en tout cas, le symbole d'une évolution notable du système de communication scientifique.

Conclusion

Le CERN est le plus grand laboratoire mondial dans le domaine de la Physique des Hautes Energies. La masse d'informations qui y circule est considérable. Le stage que j'ai effectué au sein de son Service d'Information Scientifique a donc été pour moi une expérience très enrichissante.

La mission qui m'a été confiée d'analyser des sources documentaires en littérature grise et de les exploiter a été pour moi très intéressante. Elle impliquait, en effet, différentes tâches assez diversifiées qui m'ont permis d'avoir un panorama assez large du travail de documentaliste : l'analyse des sources, la réflexion sur la politique d'acquisition, le traitement des données. Cela m'a permis de mettre en œuvre une réflexion sur cette forme de traitement du document, et en particulier sur le programme *Uploader*.

Le fait de travailler, principalement, sur les prépublications est un des aspects positifs de ce stage. J'ai pu, en effet, m'initier à ce type de documents sur lequel je connaissais peu de choses. Il me semble par ailleurs que les *preprints* sont des éléments de plus en plus importants dans la communication scientifique. On le voit par la prolifération des serveurs de prépublication. Avec eux, le schéma traditionnel est brisé. Les *preprints* prennent, en effet, autant d'importance que les articles publiés.

Enfin, il me semble important de noter que ce processus de traitement du document implique une évolution dans le travail du documentaliste, en particulier, le travail d'analyse des sources. C'est, à mon sens, un des rôles les plus importants du documentaliste à notre époque. J'emprunte quelques mots à Maria Witt pour définir cela : « *Il se trouve que les meilleurs moteurs de recherche ne couvrent qu'un tiers du WEB, le reste restant invisible et caché dans les bases de données. De cette façon, les bibliothécaires sont utiles plus que jamais pour sélectionner, classer et fournir l'information* »¹.

Le documentaliste n'est plus chargé, comme auparavant de la conservation du savoir. Il a la redoutable tâche de diffuser et de gérer toutes ces données qui prolifèrent chaque jour, de plus en plus. Il devient une sorte d'intermédiaire entre le 'faiseur d'information' et l'utilisateur et comme ces deux rôles sont interchangeables, il les endosse tour à tour pour s'adapter à la demande qu'on lui fait.

¹ [10] Witt, Maria / *Description des documents électroniques et les besoins des utilisateurs*. Médiathèque de la cité des sciences : 2001.

Bibliographie

■ La Communication scientifique

- [1] Chaney, Eliane ; Bulliard, Catherine ; Christiansen, Caroline / *Une bibliothèque de recherche face à l'édition électronique : l'exemple du CERN. Bulletin des Bibliothèques de France*, février 1999, n° 2. [page consultée le 15 septembre 2002].
<http://www.enssib.fr/bbf/bbf-99-2/05-cressent.pdf>
- [2] Chartron, Guislaine / *La presse scientifique sur les réseaux*. 1995, Revue électronique Solaris
<http://www.Unicaen.fr/bnum/jelec/Solaris>
- [3] Laloe, Franck / *La communication directe et les archives ouvertes*. 2001 [page consultée le 15 septembre 2002].
http://www.idt.fr/idt/pages_fra/actes/actes2001/page7bis.htm
- [4] La Vega Josette, de / *La communication scientifique à l'épreuve de l'Internet : l'émergence d'un nouveau modèle*. Villeurbanne : Presses de l'ENSSIB, 2000. 253 p. ISBN 2-910227-29-4
- [5] Le Moal, Jean-Claude / *La documentation numérique - Concurrences et rivalités*, BBF 2002 Paris, t 47 n1. P68-72. [page consultée le 15 septembre 2002].
http://bbf.enssib.fr/bbf/html/2002_47_1/2002-1-p68-lemoal.xml.asp
- [6] Line, Maurice / *Accéder ou acquérir, une véritable alternative pour les bibliothèques ? Bulletin des Bibliothèques de France*, 1996. [page consultée le 15 septembre 2002].
<http://www.enssib.fr/bbf/bbf-96-1/07-line.pdf>
- [7] Pignard, Nathalie ; Paillart, Isabelle (dir.) / *Les nouvelles formes de publication scientifique sur Internet - La remise en cause du modèle éditorial traditionnel*. Juin 2000. 111 p.
Mémoire soutenu pour l'obtention du DEA en Sciences de l'Information et de la Communication, Université Stendhal, Grenoble, juin 2000.
- [8] Vicens, Jens ; Servettaz, Marie-Jeanne ; Chaney Eliane / *Une offre de services adaptée aux chercheurs*. Genève, CERN : 2001.
Consultable In *Bulletin des Bibliothèques de France*, 2001. n° 2. [page consultée le 11 avril 2002].
http://bbf.enssib.fr/bbf/html/2001_46_2/2001-2-p66-chaney.xml.asp
- [9] Volland-Nail, Patricia / *L'information scientifique et technique : nouveaux enjeux documentaires et éditoriaux*. INRA : 1997

[10] Witt, Maria / *Description des documents électroniques et les besoins des utilisateurs*. Médiathèque de la cité des sciences : 2001.

▪ Le traitement automatique de documents

[11] Cart, Catherine ; Geretschläger, Ingrid / *Automatisation du traitement des documents CERN*. Genève : CERN, 25 Jan 1999. 6 p. [page consultée le 15 septembre 2002].
<http://preprints.cern.ch/archive/electronic/cern/preprints/open/open-99-068.pdf>

[12] Collignon, Isabelle ; Geretschläger, Ingrid (dir.) / *Le traitement de la littérature grise à la bibliothèque du CERN*. Genève : CERN, 1998. 69 p.
Mémoire de DEUG-DIST : I.U.P./Univ. Lyon 1, soutenu le 31 Juillet 1998.

[13] Deroche, Catherine ; Geretschläger, Ingrid (dir.) ; Jerdelet Jocelyne (dir.) / *Automatisation partielle du traitement de la littérature grise dans le service d'information scientifique du CERN*. Genève : CERN, 1998. 59 p.
Mémoire de D.E.S.S. Sci. Inf. : ENSSIB/Univ. Lyon 1. [page consultée le 15 septembre 2002].
<http://preprints.cern.ch/archive/electronic/cern/preprints/thesis/thesis-98-019.p.s.gz>

[14] Pignard, Nathalie ; Geretschläger, Ingrid (dir.) ; Jerdelet Jocelyne (dir.) / *Le traitement informatisé des ressources électroniques - importations automatiques à l'aide du programme Uploader*. Rapport confidentiel remis au Service de l'Information du CERN, octobre 2000.

[15] Pignard, Nathalie ; Bouquillion, Philippe (dir.) / *Les bases de données scientifiques sur Internet : la comparaison de trois bases de conférences de physique avec celle du CERN*. Genève : CERN, septembre 1999.
Rapport de stage (confidentiel) dans le cadre de la Maîtrise d'Information Communication, Université Lumière Lyon 2.

[16] Pignard, Nathalie ; Geretschläger, Ingrid ; Jerdelet, Jocelyne / *Le traitement informatisé de ressources électroniques au Service de l'Information Scientifique du CERN. Le Documentaliste - Sciences de l'Information*, mars 2001, vol. 38, n°1, p. 24-34.

[17] Vesely, Martin ; Vigen, Jens (dir.) / *Using Internet/Intranet Technologies in Library Automation*. Genève : CERN, 2000. 67 p. [page consultée le 15 septembre 2002].
Mémoire de thèse : Université de Prague.

Annexes

ANNEXE 1 : Organigramme de la division ETT

ANNEXE 2 : Schéma général de l'importation avec l'utilisation de l' *Uploader*

ANNEXE 3 : Exemple de bilan d'analyse et d'exploitation de sources :
Preprint Database UMIST : bilan des analyses et de l'importation, source exploitée.

ANNEXE 4 : Exemple de bilan d'analyse et d'exploitation de sources :
Computer Science Bibliography : bilan de l'analyse, source abandonnée.

ANNEXE 5 : Exemple de bilan d'analyse et d'exploitation de sources
Pascal-ConnectSciences : bilan de l'analyse et de l'importation, source abandonnée provisoirement.

ANNEXE 6 : Tests de pertinence des mots-clefs pour définir une équation de recherche pour Pascal-ConnectSciences

ANNEXE 7 : Schéma de fonctionnement de l'*Uploader*

ANNEXE 8 : Les fichiers de données

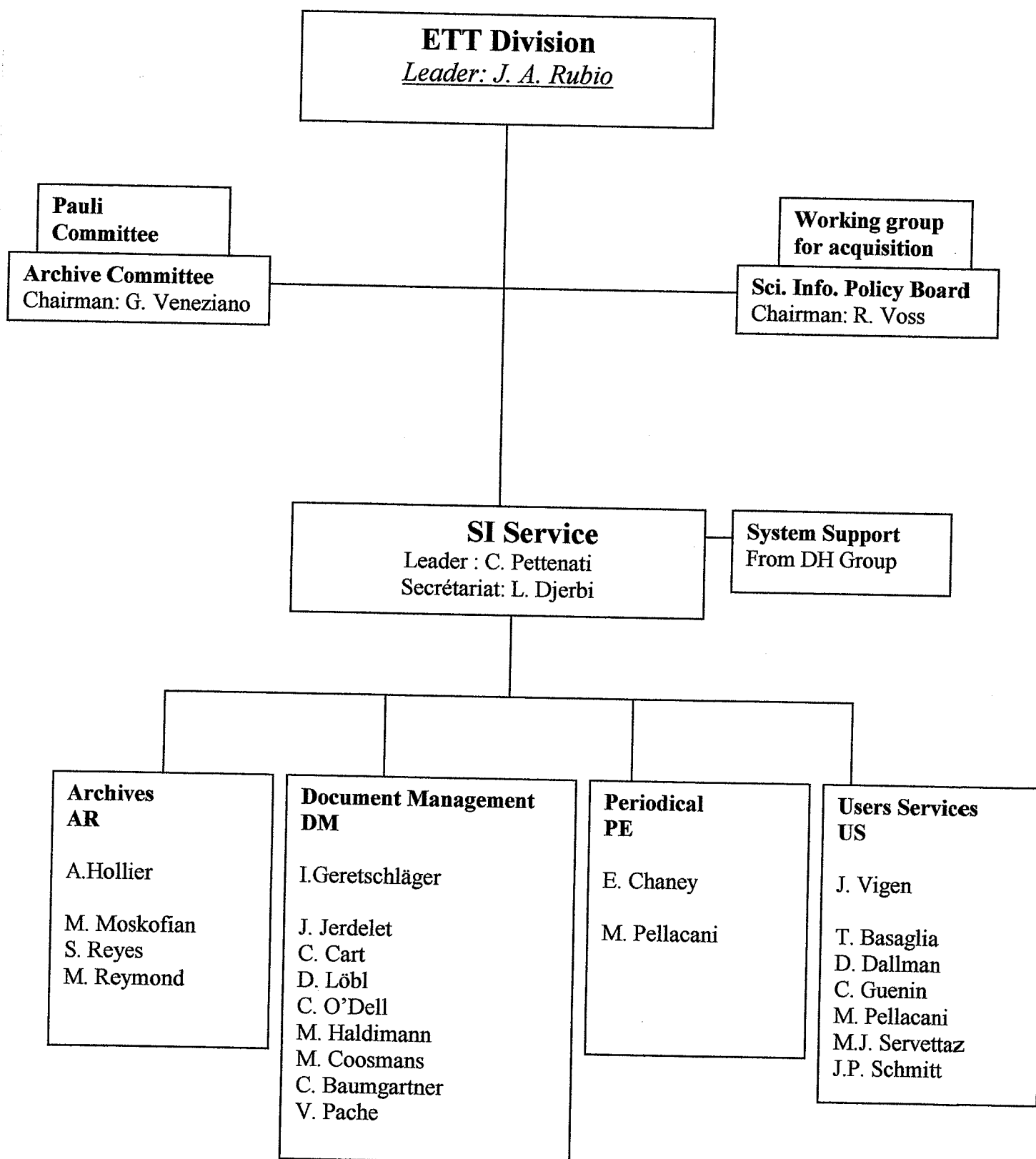
ANNEXE 9 : Exemple d'un fichier de données de l'*Umist Preprint Database*

ANNEXE 10 : Statistiques des importations manuelles et automatiques de littérature grise au CERN des années 2000-2001.

ANNEXE 11 : Liste des sources exploitées aux CERN en 2001.

ANNEXE 1

ORGANIGRAMME DE LA DIVISION ETT



ANNEXE 2

SCHEMA GENERAL DE L'IMPORTATION AVEC L'UTILISATION DE L'*UPLOADER*.

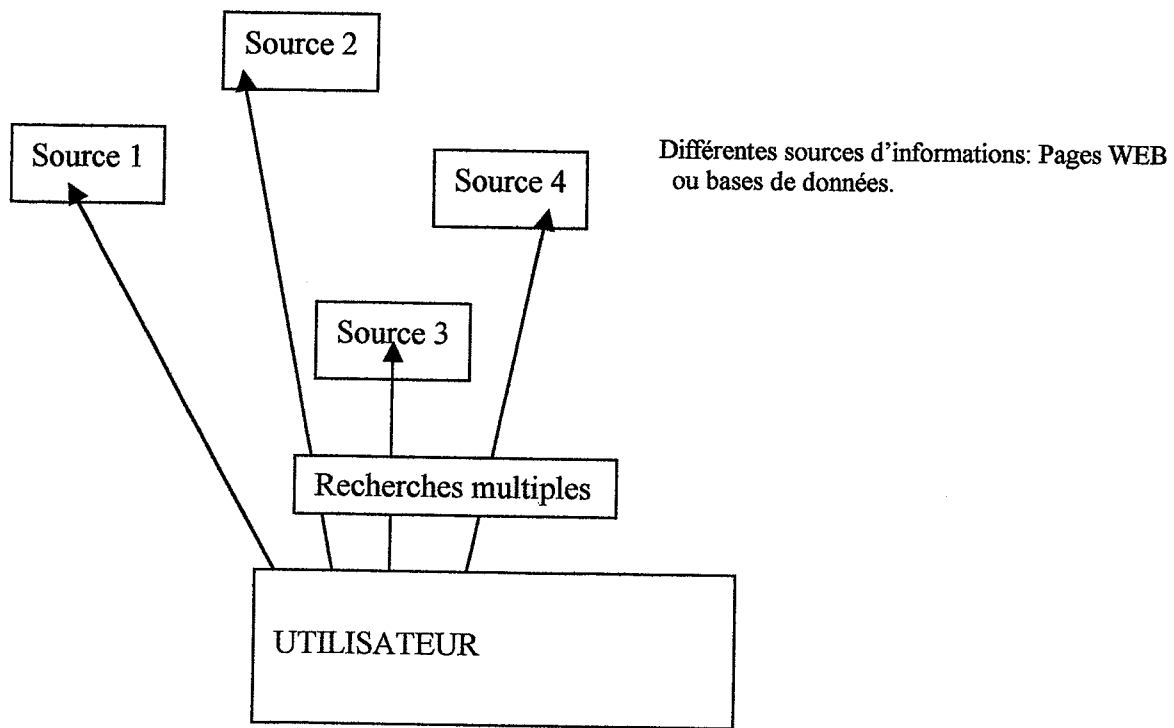
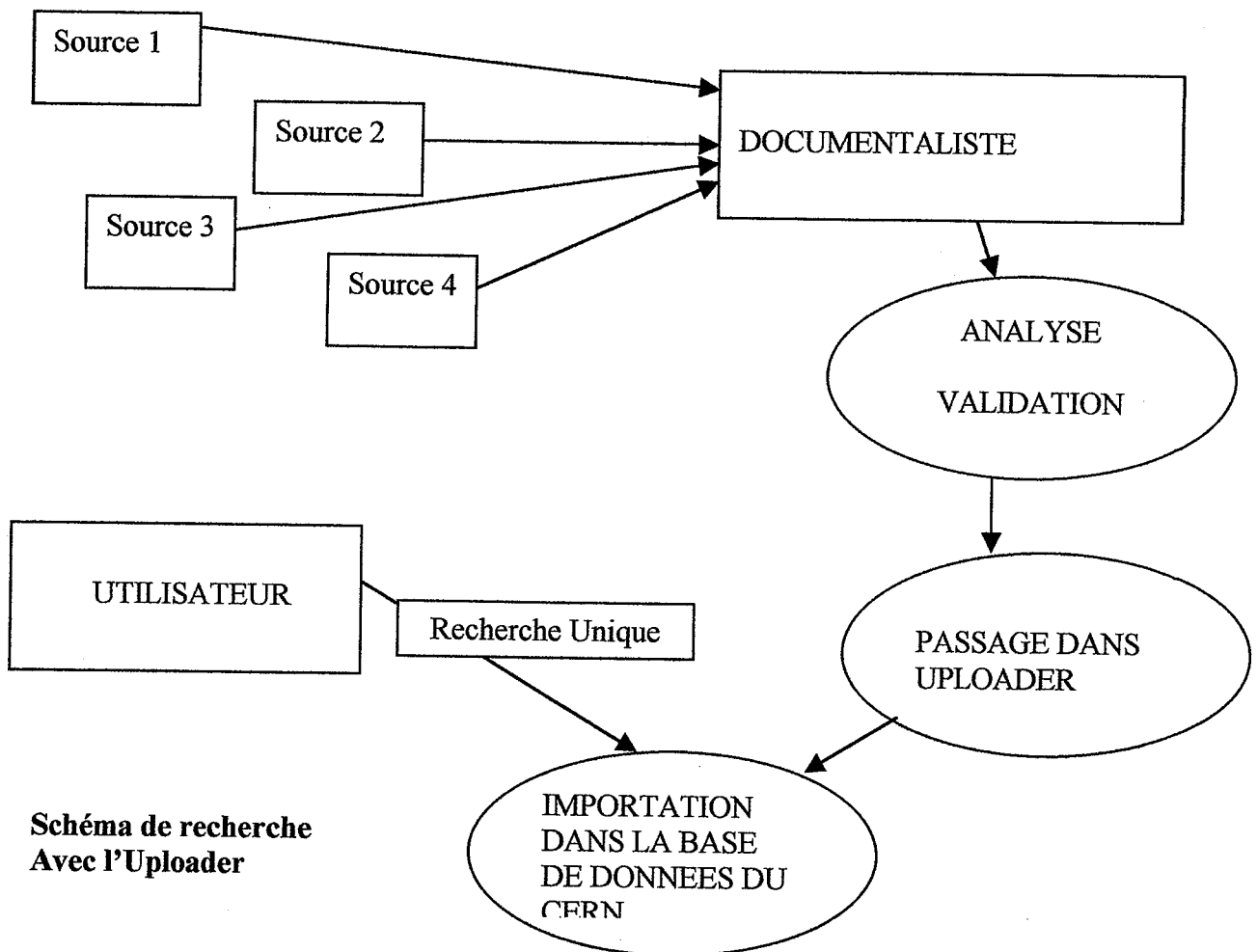


Schéma de recherche classique



ANNEXE 3

EXEMPLE DE BILAN D'ANALYSE ET D'EXPLOITATION DE SOURCES

Preprint Database UMIST : bilan des analyses et de l'importation, source exploitée.

Preprint Database System

<http://www.phy.umist.ac.uk/research/preprint.shtml>

Présentation de la base

Il s'agit de la base de données de Preprint de l'UMIST (the University of Manchester Institute of Science and Technology). La période couverte s'étend de 1995 à l'année en cours.

Les notices contiennent parfois un lien vers le texte intégral, accessible gratuitement.

Site consulté en avril 2002 : 273 documents

Les documents sont divisés en 5 groupes thématiques, ils correspondent au groupe de recherche de l'UMIST :

- AstroPhysics
- Atmospheric Physics
- Condensed Matter
- Plasma Physics
- Theoretical Physics

La recherche

Il s'agit d'une recherche combinée, on peut remplir plusieurs champs :

- On peut choisir le groupe thématique, mais on peut n'en sélectionner qu'un à la fois. (Menu déroulant)
- Champ : auteur (100 caractères maximum)
- Champ : titre (100 caractères maximum)
- Bouton radio : Publié ou non-publié
- Champs : année

Caractéristiques de cet outils de recherche :

- ✓ La casse compte

Exemple : Si on recherche le terme « chemistry » dans un titre, et si on écrit sans C majuscule, les résultats ne comprendront que les titres contenant « chemistry ».

En analysant les notices, nous avons constaté que les titres n'étaient pas tous écrits de manière uniforme, parfois, il y a des majuscules à tous les mots, parfois non.

Il est donc nécessaire, pour faire une recherche efficace de chercher avec et sans majuscule.

- ✓ La requête peut s'effectuer sur une partie de mot seulement, l'outil de recherche est capable de faire des suppositions sur la fin, ou le début d'un mot. En revanche, il n'y a pas d'approximation s'il y a une faute d'orthographe dans le nom.

Exemple : « chemis »----- « Astrochemistry » -----« chemistry »

- ✓ On peut utiliser les opérateurs booléens, mais il y a seulement « and » et « or », de plus, ce sont des boutons-radio et on ne peut les inclure directement dans une équation de recherche.

Structure et champs des notices

Les champs proposés sont:

- Numéro d'identification
- Auteur
- Titre
- Lien vers le texte intégral
- Publication
- Lien vers notice plus complète

Remarques: Les informations de publication ne sont pas toujours très bien structurées
Il n'y a pas systématiquement un lien vers le texte intégral.

Exemple:

Preprint: UMIST:Phys:Ast:0-1
 Authors: T J Millar
 Title: Molecules in High-Mass Star-Forming Regions - Theory and Observation
 Text at this url
 Published in In 'Science with the Atacama Large Millimeter Array'ASP Conference Series, A
 Wootten, ed
 Detailed view/Edit

La configuration

Le nom des fichiers de configuration commence par **UMIST**

Utilisation de KB :

UMISTan.kb : utilisé pour transformer les années au format voulu

UMISTsu.kb : utilisé pour indiquer la bonne catégorie dans le champ SU

Knowledge_base : pour transcrire à la forme voulue les titres de journaux.

Particularités :

- Importation en semaine courante
- Champs NI=Liste des auteurs pour permettre vérification
- Les notices vont dans la BASE 13 (preprints publiés avec toutes les informations sur le publication) ou la BASE 11 (Notices des preprints non publiés ou des preprints publiés dont les informations de publication ne sont pas complètes)

L'importation

- *La recherche*

Première importation (Mai 2002) : environ 105 notices

- Aller sur page d'accueil de la base :
<http://www.phy.umist.ac.uk/research/preprint.shtml>
- Cliquer sur SEARCH (<http://www.phy.umist.ac.uk/cgi-bin/zyview/D=preprint/V=search>)
- Chercher par mots du titre en ne sélectionnant aucun sujet
- Sélectionner <OR>
- Sélectionner : 60
- Enumérer liste de mots-clés (intéressants pour le CERN):

Quantum Cluster Gauge Many Body Hamiltonian Phase Transition particle Nucl Monte
COMPASS Boson Spin

Remarque : cette recherche a été faite pour la première importation. Pour les années suivantes, il faudra faire une recherche par sujet et par année .

Les années suivantes (Importation annuelle)

- Cliquer sur SEARCH
- Sélectionner sujet (On ne peut interroger qu'un sujet à la fois)

Les sujets importés seront :

- AstroPhysics
- Condensed Matter
- Plasma Physics
- Theoretical Physics
- Dans le champs « DATE » inscrire la date de l'année précédent l'année en cours.
- Vérifier que l'opérateur « AND » est sélectionné.
- Sélectionner « 60 » pour le nombre de résultats par page.

NB : Faire attention aux minuscules et majuscules

- *La récupération des données*
- Prendre le code source HTML. Chaque notice est entourée par les balises <BI> et </BI>. **Ne prendre que les notices**, éliminer l'en-tête du fichier HTML ainsi que ce qui se trouve à la fin.
- Copier le code source dans un fichier emacs, enregistrer ce fichier de façon à ce qu'il ait un nom unique. Faire autant de fichier qu'il y a de page HTML.

- *Séparation des données*

</L>

- *Vérification/modification avant le passage dans l'Uploader*

- Les auteurs sont inscrits ainsi :

auteur1, auteur2, auteur3 and auteur4

supprimer tous les « and » et les « & » du champs auteur et les remplacer par une virgule suivie d'un espace :

auteur1, auteur2, auteur3, auteur4

- Vérifier les champs qui contiennent les informations sur la publication :
S'il s'agit d'un journal publié : les informations doivent être disposées de cette façon :

Titre de la publication, volume (Année) pages

NB : Faire très attention aux virgules, espaces parenthèses.

- S'il s'agit de conférences, il n'y a pas de champs PR, les informations vont dans le champ LKR, dans ce cas, il n'y a pas besoin de faire des modifications.
- Supprimer toutes les notices de livres édités : il sont généralement notés avec une indication avant le titre : **[edited book]**

- *Vérification après le passage dans Uploader*

- Vérifier le champs PR, dans le cas des journaux publiés : il doit se présenter ainsi :

\$\$p Titre du journal \$\$v Volume \$\$y Année de publication \$\$c Pages

S'assurer que la notice est en BASE 13.

- Si dans la notice initiale, il n'y avait pas toutes les informations, il est possible que le champ PR ne soit pas correctement écrit, il faut alors enlever les sous-champs superflus. Vérifier que la notice se trouve en BASE 11
- Si il n'y a pas de publication, vérifier que la notice est en BASE 11.
- Vérifier que les auteurs ainsi que leurs affiliations sont correctement retranscrits grâce au champs NI.

ANNEXE 4

EXEMPLE DE BILAN D'ANALYSE ET D'EXPLOITATION DE SOURCES

Computer Science Bibliography : bilan de l'analyse, source abandonnée

Computer Science Bibliography
<http://www.informatik.uni-trier.de/~ley/db/index.html>

Ce site se trouve au sein de l'université Trier.

Il s'agit d'une bibliographie sur les sciences de l'informatique. On trouve plusieurs

entrées:

- Conferences
- Journals
- Series
- Books

Ici, ce sont les conférences qui nous intéressent.

On trouve, dans un premier temps, les notices des conférences :

Par exemple :

18. AAAI / 14. IAAI 2002: Edmonton, Alberta, Canada
AAAI/IAAI 2002 Home Page
(in 2001 IJCAI was in North America)

17. AAAI / 12. IAAI 2000: Austin, TX
Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth
Conference on Innovative Applications of Artificial Intelligence, July 30
- August 3, 2000, Austin, Texas, USA. AAAI Press / The MIT Press, 2000, ISBN 0-262-
51112-6
Contents - AAAI/IAAI 2000 Home Page

....

Les notices sont assez complètes. Les différents champs sont clairement séparés.

Indications présentes :

Le lieu, le nom de la conférence, sa date et s'il y a lieu, la mention de publication et son ISBN.

Pour chaque conférence, on a un lien vers la liste des *proceedings* :

Ex :

Invited Papers

Umeshwar Dayal, Meichun Hsu, Rivka Ladin:
Business Process Coordination: State of the Art, Trends, and Open Issues. 3-13
Electronic Edition (link)

Egbert-Jan Sol:
Ambient Intelligence with the Ubiquitous Network, the Embedded Computer Devices
and the Hidden Databases (abstract). 14
Electronic Edition (link)

Les notice de *proceeding* sont toujours rédigées de la même manière :

Prénom Nom, Prénom Nom, ... : Titre : Sous titre. Pages Lien vers le texte intégral
--

NB :

- Il n'y a pas toujours un lien vers le Full Text
 - les liens ne sont pas toujours accessibles librement, parfois l'accès est refusé : Le Full Text n'est pas toujours libre de droit. Ex : ACM, Springer
- Il faut obtenir des autorisations.

Sujets Traités Dans Cette Bibliographie :

Algorithms –
Artificial Intelligence
Cryptology/Security
Database Systems –
Digital Libraries –
Distributed Computing
Hypertext
Information Retrieval
Logic Programming
Machine Learning - Multimedia
Operating Systems
Theoretical Computer Science

⇒ Ces sujets peuvent être intéressants pour la base du CERN, qui possède une rubrique sujet « Computer/Computing ».

La recherche

Il y a plusieurs possibilités de recherches :

➤ Recherche par nom d'auteur : un seul champ, peu intéressant.
<http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/index.html>

➤ Recherche par titre : Idem
<http://www.informatik.uni-trier.de/~ley/db/indices/t-form.html>

➤ Recherche avancée :
<http://www.informatik.uni-trier.de/~ley/db/indices/query.html>

Plusieurs champs sont disponibles : possibilité de combiner l'auteur et le titre, année, page. Si on cherche une conférence, il y a un champ spécial pour inscrire le titre. Cependant, il est nécessaire de connaître le nom de la conférence. On ne peut, par exemple, pas faire la requête : toutes les conférences de 2002.

Le moteur est capable de faire des suppositions sur la fin d'un mot et de chercher tous les mots qui contiennent l'élément indiqué.

Les caractères spéciaux renvoient aux caractères simples correspondants.

Cependant, si le caractère est nommé en code HTML, c'est le caractère spécifié qui est recherché.

Source où il est difficile de faire une équation de recherche adaptée à notre objectif.

- Autre possibilité pour la recherche : on peut entrer dans la base grâce à une entrée « subject ».

<http://www.informatik.uni-trier.de/~ley/db/subjects.html>

Algorithms

Artificial Intelligence

Cryptology/Security

Database Systems –

Digital Libraries –

Distributed Computing

Hypertext

Information Retrieval

Logic Programming

Machine Learning - Multimedia

Operating Systems

Theoretical Computer Science

Pour chaque sujet, on a la liste des conférences qui lui sont liées. Cela permet de cibler un peu la recherche.

NB : Certaines conférences ne sont référencées sous aucun sujet. La recherche n'est donc, par définition pas exhaustive.

Pour conclure, il faut dire que cette source peut être très intéressante pour enrichir la base du CERN, cependant, une importation totalement automatique est impossible dans la mesure où on ne peut faire une équation de recherche adéquate.

On peut toutefois envisager d'exploiter la base, en important les notices des *proceedings* des conférences que l'on possède déjà au CERN. Cela pourrait se faire par une analyse manuelle des différentes conférences par sujet.

On pourrait se limiter aux 3 dernières années puis effectuer une mise à jour annuelle.

ANNEXE 5

EXEMPLE DE BILAN D'ANALYSE ET D'EXPLOITATION DE SOURCES

Pascal-ConnectSciences : bilan de l'analyse et de l'importation, source abandonnée provisoirement.

CONNECTSCIENCES :
Analyse du portail mai-juin 2002
<http://connectsciences.inist.fr/>

ConnectScience est un portail du CNRS qui permet un accès à plusieurs sources d'information scientifique :

- Catalogue d'articles (accès au catalogue des articles issus du fonds INIST)
- Bibliosciences (accès multibases réservé aux unités CNRS),
- Nouveautés (trois mois) de la base PASCAL (interrogeable via LexiQuest Mine®, outil d'aide à la formulation de requête),
- Nouveautés de la base FRANCIS (interrogeable via LexiQuest Mine®, outil d'aide à la formulation de requête),
- Base THESA (base des thèses en cours dans les grandes écoles). : THESA signale les sujets (titres) de thèses en cours, dès la première année d'inscription en thèse, du candidat au doctorat jusqu'à la soutenance de la thèse. Le signalement de la thèse est conservé dans la base un an après la délivrance du diplôme ; la référence bibliographique complète de la thèse figurera alors dans la base de données bibliographiques TheseNet, qui signale les thèses soutenues en France dans les Universités et les Grandes Ecoles habilitées.

Pour accéder aux ressources, il est nécessaire de s'inscrire, cette inscription est gratuite. Il faut un nom d'utilisateur et un mot de passe.

La base INIST possède environ 7 millions de références, provenant de 9000 périodiques français et internationaux.

Par l'intermédiaire de ConnectScience, nous n'avons accès qu'aux notices. Pour avoir le texte intégral, il est nécessaire de le commander à l'INIST, ce service est payant. 50% des notices sont cependant dotées d'un abstract.

L'accès à PASCAL par ConnectScience :

➤ Type de documents :

- Articles de périodiques
- Thèses
- Congrès

➤ Période couverte : les trois derniers mois.

➤ Outils de recherche :

- Recherche simple (Cf Catalogue d'articles)
- Recherche avancée : Lexiquest Mine

Possibilité de sélectionner un thème pour affiner la recherche, liste de différents thèmes de la base : Biologie, Chimie, Mathématique/Physique, Médecine, Science de l'information, Science de l'ingénieur, Science de l'univers. Cette fonctionnalité n'est pas encore vérifiée.

- On propose aussi une liste de concepts pour chaque thème, ces concepts sont définis par rapport aux documents de la base PASCAL de l'année en cours.
- Equation de recherche : opérateurs booléens [AND] [OR] [NOT], possibilité d'inscrire de nombreux termes dans l'équation.

- Autres paramètres à prendre en compte : on peut choisir sur quel type d'information porte la requête (cases à cocher) : Vocabulaire contrôlé, Auteurs, organisme, Sources, Lieu, terme. NB : Cette fonction n'est, pour l'instant pas effective.
- Recherche sur le mot précis, ou sur une partie de mot. Ex : si on inscrit CERN, le moteur cherche seulement ce mot là, si on inscrit CERN*, il va chercher tous les mots qui contiennent cette chaîne de caractère. NB : Cette fonctionnalité n'est pas complètement fiable.

➤ Notice :

Données de la première notice : titre

Auteur(s)
Source : titre volume numéro pages
Date de publication
Pays : ETATS-UNIS
Langue du texte
Type de document
Numéro INIST

Données de la notice complète : Auteur(s)

Affiliation(s)
Source
ISSN
Date de publication
Pays
Langue du texte
Type de document
Résumé :
Code de classement
Mots-clés français
Mots-clés anglais
Localisation
Numéro INIST

➤ Récupération des données :

- Possibilité d'enregistrer les résultats de recherche dans des dossiers, ce qui permet de combiner les recherches et dédoubler les éléments.
- On peut aussi enregistrer les données en format texte, mais cette fonction ne marche pas correctement pour les articles de la base PASCAL.
- On peut accéder au code source de la page HTML qui affiche les résultats.
- La fonction « alerte » fonctionne sur cette base, seulement, on n'a accès qu'à un lien sur les notices, nous n'avons pas directement toutes les informations, de ce fait, il est impossible d'importer automatiquement le code-source.
- fichier texte.

Intérêt de cette source :

Les documents que l'on trouve sont surtout des articles de périodiques, cela peut être intéressant pour compléter les informations des notices de nos Preprints. En effet, il y a peu de documents que nous ne possédions déjà au CERN.

Par exemple : pour une recherche « LHC and Large Hadron Collider » nous avons 13 résultats dans la base PASCAL

- 9 enregistrements sont déjà dans la base du CERN en base 11
- 3 enregistrements sont déjà dans la base du CERN avec toutes les informations
- 1 est inconnu

Autre exemple : « LHC and Physics » : 44 résultats (recherche effectuée début juin 2002 dans la base PASCAL)

- BASE 11 : 26 résultats
- BASE 13 : 2
- BASE 14 : 1
- Non connu dans la base : 15 (6 trouvés dans INSPEC, 9 non trouvés)

En conclusion, on peut induire le fait que les notices que nous n'avons pas dans la base seront importées grâce à INSPEC.

La fonction UPLOAD n'est donc pas nécessaire, la base PASCAL doit davantage servir à une mise à jour (UPDATE) de la BASE 11.

En revanche, il peut être intéressant d'importer les thèses.

ESSAI D'IMPORTATION A PARTIR DE LA BASE PASCAL

Juin 2002

CONFIGURATION :

Nom de la configuration : COSCI :

Particularités de la configuration :

- Séparateur de notices : '<valign= "middle">'
- La recherche se fait à partir des 4 premiers mots du titre et sur le nom de l'auteur.
- Utilisation de la knowledge_base pour que les titres des périodiques soient correctement présentés.
- Matching = 1
- Pas de Upload

TEST D'IMPORTATION

Comme le système d'alerte ne permet pas une récupération correcte des données, on a utilisé le code source HTML.

➤ Equation de recherche :

Dans le mode recherche experte, on sélectionne, par un menu déroulant, la partie Physique/Mathématique de la base.

(CERN OR LEP OR DELPHI OR CMS OR ALEPH OR ALICE OR CHARM OR CLIC OR COMPASS OR CPLEAR OR FELIX OR ISOLDE OR LARGE HADRON COLLIDER OR LHC OR LHCb OR NOMAD OR OBELIX OR REX OR SPIN MUON OR TILECAL OR TOSCA OR TOTEM OR OPAL OR PROTON SYNCHROTON) NOT ((CERN OR LEP OR DELPHI OR CMS OR ALEPH OR ALICE OR CHARM OR CLIC OR COMPASS OR CPLEAR OR FELIX OR ISOLDE OR LARGE HADRON COLLIDER OR LHC OR LHCb OR NOMAD OR OBELIX OR REX OR SPIN MUON OR TILECAL OR TOSCA OR TOTEM OR OPAL OR PROTON SYNCHROTON) AND (THESE OR THESIS))

NB : Possibilité d'affiner la recherche en ajoutant des termes plus généraux tels que Physique, Nucléaire...

- Les résultats s'affichent sur des pages HTML. Il y a seulement 10 résultats par page.
- Prendre dans le code-source (Faire un copier/coller dans un fichier emacs, ou enregistrer la page en fichier texte). Veiller à ne prendre que ce qui concerne les notices, effacer le reste :

Entre le premier (compris) et la fin de </TABLE>
 /\ /\ /\ /\ /\ /\ /\ /\ /\

- Passage dans l'uploader, pas besoin de vérification préalable.
- Après le passage dans l'uploader, vérifier les champs PR, quand les titres des périodiques sont en majuscule, ils ne se trouvent pas dans la knowledge_base, il faut alors rechercher la forme correcte.

BILAN DE LA PREMIERE IMPORTATION : 10 juin 2002

788 résultats pour la recherche d'articles.

Une fois les fichiers passés dans l'UPLOADER, la taille des fichiers est beaucoup plus petite. Une centaine de notices a été mise à jour (118 exactement): cela est assez faible par rapport au nombre d'entrées.

Cela signifie qu'un certain nombre d'articles ne sont pas connus dans la base, deux explications possibles :

- la recherche est peu fiable, par conséquent, il est normal que nous ne possédions pas les articles trouvés (autres domaines que celui qui nous intéresse) .
- Il aurait fallu faire une importation et non une mise à jour.

Test de fiabilité de la recherche :

Pour 388 notices issues de la recherche montrée ci-dessus, une analyse rapide a été faite pour vérifier la pertinence de leur présence.

(Analyse faite par le titre, recherche des mots-clés de l'équation)

- 90 jugées pertinentes
- 298 jugées non-pertinentes.

Le moteur de recherche de la base PASCAL n'est donc pas très performant.

BILAN GENERAL

- ✓ Grande quantité d'informations accessible par ce portail mais les données ne sont pas complètes : pas d'accès direct aux notices complètes, pas de texte intégral gratuit, l'abstract n'est pas obligatoirement présent.
- ✓ Diversité de ces données, par conséquent la spécialisation est limitée
- ✓ Difficulté d'affiner les recherches dans le domaine désiré
- ✓ Certaines options proposées sont encore peu fiables

Pour l'instant, la base PASCAL n'est pas assez fiable pour être utilisée de façon efficace.

COMPLEMENT D'ANALYSE SUR LE PORTAIL CONNECTSCIENCE

Fin juillet 2002

Compléments sur plusieurs points qui n'ont pas été traités lors de l'analyse précédente, la base évolue.

❖ Les termes d'indexation INIST

On peut effectuer la recherche grâce aux termes d'indexation INIST : il suffit d'écrire le mot-clé que l'on souhaite et le moteur nous donne tous les termes déjà répertoriés dans le catalogue qui s'en approchent.

Par exemple : si on tape NUCLEAIRE, on obtient : énergie nucléaire, nucléon, physique nucléaire...

➤ Cependant, par ce moyen là non plus la recherche n'est pas toujours pertinente. Par exemple : le mot-clé SPIN → on obtient 4062 résultats. (que l'on sélectionne la partie mathématique/Physique ou non).

En regardant les résultats, on trouve des données non pertinentes par rapport aux critères de la recherche.

En revanche, pour les mots-clés assez précis comme SPIN NUCLEAIRE : on a une homogénéité des résultats → 54 résultats pertinents.

Comme on l'a déjà vu, la recherche peut se faire en anglais et en français, les termes d'indexation peuvent se faire dans les deux langues. On remarque cependant que lorsqu'on utilise les mots-clés anglais, la recherche ne fonctionne pas : ERROR 404

→ Peut-être, pourrait-on essayer de chercher une équation de recherche à partir des termes d'indexation.

Seulement, ils sont très nombreux, et très précis. Nous avons vu qu'il était peu prudent d'utiliser les termes trop généraux (recherche peu fiable). Il faudrait réunir ceux qui ont le plus d'intérêt pour nous. (Cf. liste)

❖ Alertes

- Les alertes peuvent être quotidiennes, hebdomadaires ou mensuelles.
- On peut désormais faire des équations complexes dans le champ des alertes
- Il ne peut y avoir que 100 documents à chaque fois puisqu'ils sont gardés dans des dossiers qui ne peuvent contenir plus. Les données sont écrasées par l'alerte suivante.
- En ce qui concerne l'accès aux données des alertes, il n'y a toujours que le titre sur la page de résultats. Lorsqu'on enregistre cette page en fichier texte, les notices issues de Pascal n'apparaissent pas = impossibilité pour l'instant de récupérer les données.

❖ Autres remarques

- Obligation d'importation par les pages de résultats, il y a seulement 10 résultats par page, si la quantité de données est importante, cela représente un temps important pour récupérer le code-source

Par exemple : S'il y a 1000 résultats, il faut récupérer le code-source de 100 pages HTML

- Accès aux données : Environnement PC, on peut enregistrer en fichier texte et avoir la notice sans les balises HTML.

Le fichier de configuration est fait, lui, à partir du code source, car en environnement UNIX, lorsqu'on enregistre en fichier texte, le code HTML apparaît.

❖ Exemples d'équation de recherche à partir de termes d'indexation

muon OR spin nucléaire OR neutrino OR boson OR méthode monte carlo OR transition de phase OR hamiltonien OR physique atomique OR physique haute énergie OR physique nucléaire OR particule matière particulaire OR haute énergie OR énergie nucléaire OR accélérateur particule OR accélérateur linéaire OR accélérateur électron OR accélérateur proton OR collision particule OR énergie collision OR accélérateur collision faisceau OR nucléon
= 1713 résultats

muon OR spin nucléaire OR neutrino OR boson OR méthode monte carlo OR transition de phase OR hamiltonien OR physique atomique OR physique haute énergie OR physique nucléaire OR particule matière particulaire OR haute énergie OR énergie nucléaire OR accélérateur particule OR accélérateur linéaire OR accélérateur électron OR accélérateur proton OR collision particule OR énergie collision OR accélérateur collision faisceau OR nucléon OR hadron OR méson OR kaon OR lepton OR large hadron collider
= 2049 résultats

➤ Bilan

Il faudrait analyser les différents termes d'indexation proposés et en faire une équation de recherche à partir de laquelle il faudrait faire une alerte. On pourrait ainsi récupérer tous les mois les notices.

Cependant, il est nécessaire d'attendre encore. L'interface qui donne accès aux Nouveautés Pascal n'est pas encore tout à fait performante. Il faut attendre qu'elle s'améliore. On peut remarquer que pendant le temps qui a séparé les deux analyses, des améliorations ont été faites, cela laisse imaginer que l'on pourra dans un futur assez proche utiliser cette source correctement.

ANNEXE 6

TESTS DE PERTINENCE DES MOTS-CLEFS POUR DEFINIR UNE EQUATION DE RECHERCHE POUR PASCAL-CONNECTSCIENCES

MOT CLE	Resultats	Pertinence des resultats
CERN	144	peu pertinent
LEP	36	peu pertinent
LEP AND CERN	10	pertinent
DELPHI	30	peu pertinent
DELPHI AND CERN	1	pertinent
DELPHI AND LEP	3	pertinent
CMS	26	peu pertinent
CMS AND CERN	2	pertinent
ALEPH	4	pertinent
ALICE	110	peu pertinent
ATLAS	158	peu pertinent
ATLAS AND CERN	3	pertinent
CHARM	31	peu pertinent
CHARM AND CERN	4	pertinent
CHORUS	9	peu pertinent
CHORUS AND CERN	0	
CLIC	6	peu pertinent
COMPASS	30	peu pertinent
COMPASS AND CERN	1	pertinent
CLEAR	0	
FELIX	153	peu pertinent
ISOLDE	14	pertinent
LARGE ELECTRON COLLIDER	28	pertinent
LHC	30	pertinent
LHC AND LARGE...	13	pertinent
LHCB	2	Non pertinent
NOMAD	14	Non pertinent
OBELIX	3	pertinent
REX	41	peu pertinent
REX AND CERN	3	pertinent
ROSE	1308	Non pertinent
ROSE AND CERN	0	
SPIN MUON	26	pertinent
TILECAL	0	
TOSCA	10	Non pertinent
TOTEM	2	Non pertinent
OPAL	52	peu pertinent
PROTON SYNCHROTON	1	pertinent
PS	505	peu pertinent
RD	578	peu pertinent
NA	11017	Non pertinent
WA	5	Non pertinent

ANNEXE 7

SCHEMA DE FONCTIONNEMENT DE L'*UPLOADER*

La transformation d'un champ « Auteur » grâce aux trois fichiers de configuration

Champ Auteur dans le code source de la notice originale :

Authors: Isabelle Cherchneff

Fichier *.extract : description de la notice originale pour en extraire les champs

Le champ que l'on veut extraire

AU---Authors: ---

Limites du champ

Fichier *.tpl :
Encodage des champs

AU---<:PRENOM:> <:NOM:>

Deux sous-champs

Fichier *.aleph.tpl

AU2::RANGE(1,1)---CER <:SYSNO:> AU2 L <:AU*::NOM:CAP::SHAPE:>,

Nom du champ
Dans Aleph

Numéro de système
attribué par l'Uploader

Récupération ds champs: on prend le sous champ
NOM du champ AU de la notice originale. Commande
Pour mettre la première lettre en majuscule (CAP)
Commande pour effacer les espaces superflus (SHAPE)

<:AU*::PRENOM::ABR(1,)::REP(-,)::SHAPE:>

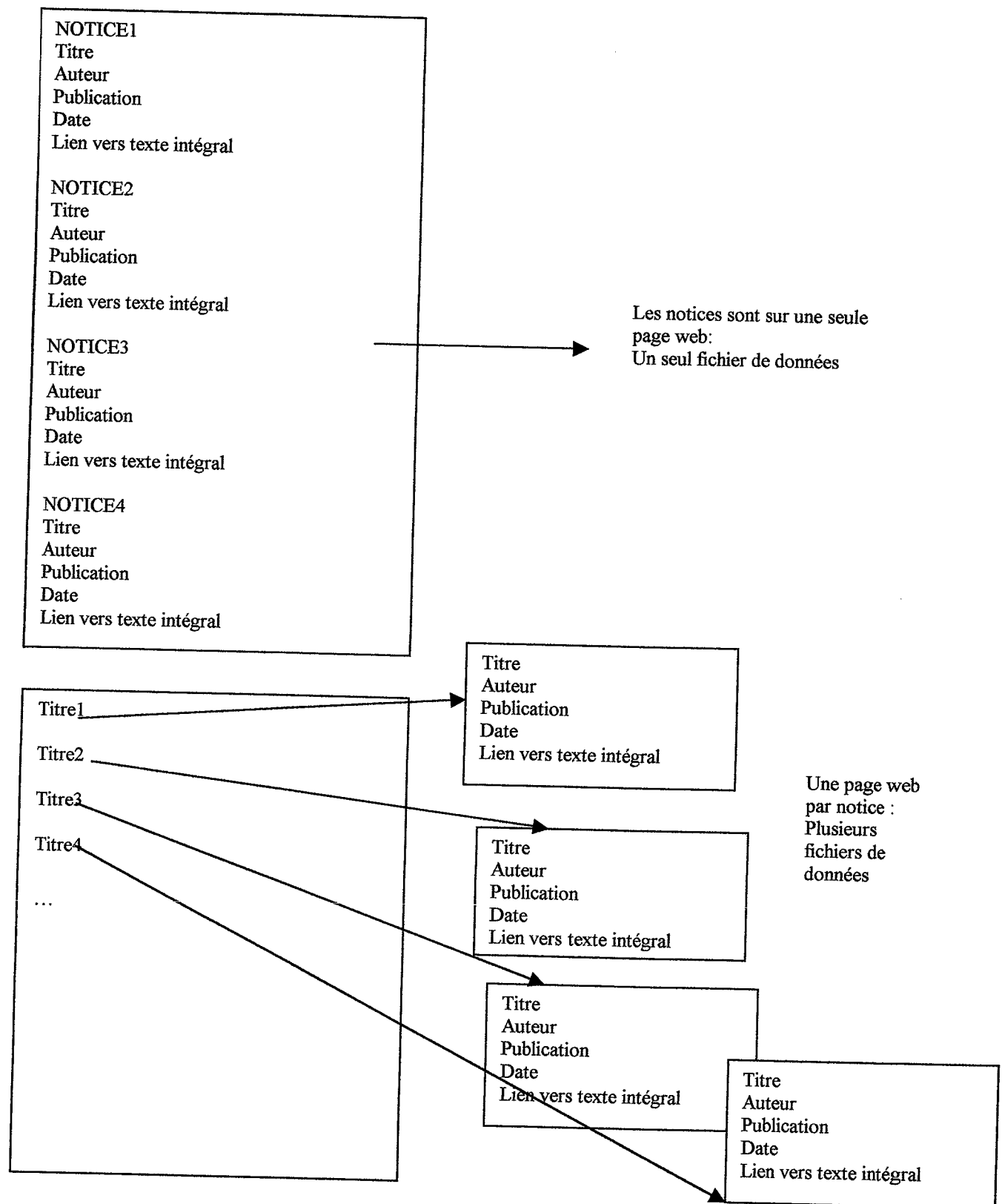
Même opération pour le sous champ PRENOM. La commande ABR l'abrège, les tirets sont supprimés :
REP(-,).

Champ Auteur dans sa forme finale à la sortie du programme Uploader:

CER 1234567 AU L Cherchneff, I

ANNEXE 8

LES FICHIERS DE DONNEES



ANNEXE 9

EXEMPLE D'UN FICHIER DE DONNEES DE *L'UMIST PREPRINT DATABASE*

Preprint Database Search Results

1. Preprint: UMIST:Phys:Ast:0-1
Authors: T J Millar
Title: Molecules in High-Mass Star-Forming Regions - Theory and Observation
Text at [this url](#)
Published in In 'Science with the Atacama Large Millimeter Array'
ASP Conference Series, A Wootten, ed.
[Detailed view/Edit](#)
2. Preprint: UMIST:Phys:Ast:0-10
Authors: H Roberts and T J Millar
Title: Modelling of Deuterium Chemistry and its Application to Molecular Clouds
Text at [this url](#)
Published in Astron. Astrophys, 361, 388-398 (2000)
[Detailed view/Edit](#)
3. Preprint: UMIST:Phys:Ast:0-11
Authors: T J Millar
Title: Interstellar and Circumstellar Chemistry: The UMIST Database
Text at [this url](#)
Published in in '2nd International Conference on Atomic and Molecular Data and their Applications'
(eds. K A Berrington & K L Bell), AIP Conference Series No. 543, pp. 81-91 (2000)
[Detailed view/Edit](#)
4. Preprint: UMIST:Phys:Ast:0-12
Authors: D Duari and J Hatchell
Title: HCN in the Inner Envelope of Chi Cygni
Published in Astronomy and Astrophysics, 358, L25-L28 (2000)
[Detailed view/Edit](#)

[La visualisation des notices sur la page WEB.](#)

```

<OL>
<LI>
Preprint: UMIST:Phys:Ast:0-1<BR>
Authors: T J Millar
<BR>
Title: Molecules in High-Mass Star-Forming Regions - Theory and Observation
<BR>
Text at <A HREF="http://saturn.phy.umist.ac.uk:8000/~tjm/alma.ps.gz"> this url
</A><BR>
Published in In 'Science with the Atacama Large Millimeter Array' <br> ASP Conference
Series, A Wootten, ed.<BR>
<A HREF="/cgi-bin/zyview/R=UMIST:Phys:Ast:0-1/D=preprint/V=shv">Detailed
view/Edit</A>
</LI>

<LI>
Preprint: UMIST:Phys:Ast:0-10<BR>
Authors: H Roberts and T J Millar
<BR>
Title: Modelling of Deuterium Chemistry and its Application to Molecular Clouds
<BR>
Text at <A HREF="http://saturn.phy.umist.ac.uk:8000/~tjm/deut1.ps.gz"> this url
</A><BR>
Published in Astron. Astrophys, 361, 388-398 (2000)<BR>
<A HREF="/cgi-bin/zyview/R=UMIST:Phys:Ast:0-10/D=preprint/V=shv">Detailed
view/Edit</A>
</LI>

```

Le code source de la page WEB.

On le copie dans un fichier texte, c'est ce fichier que l'on passe dans l'Uploader

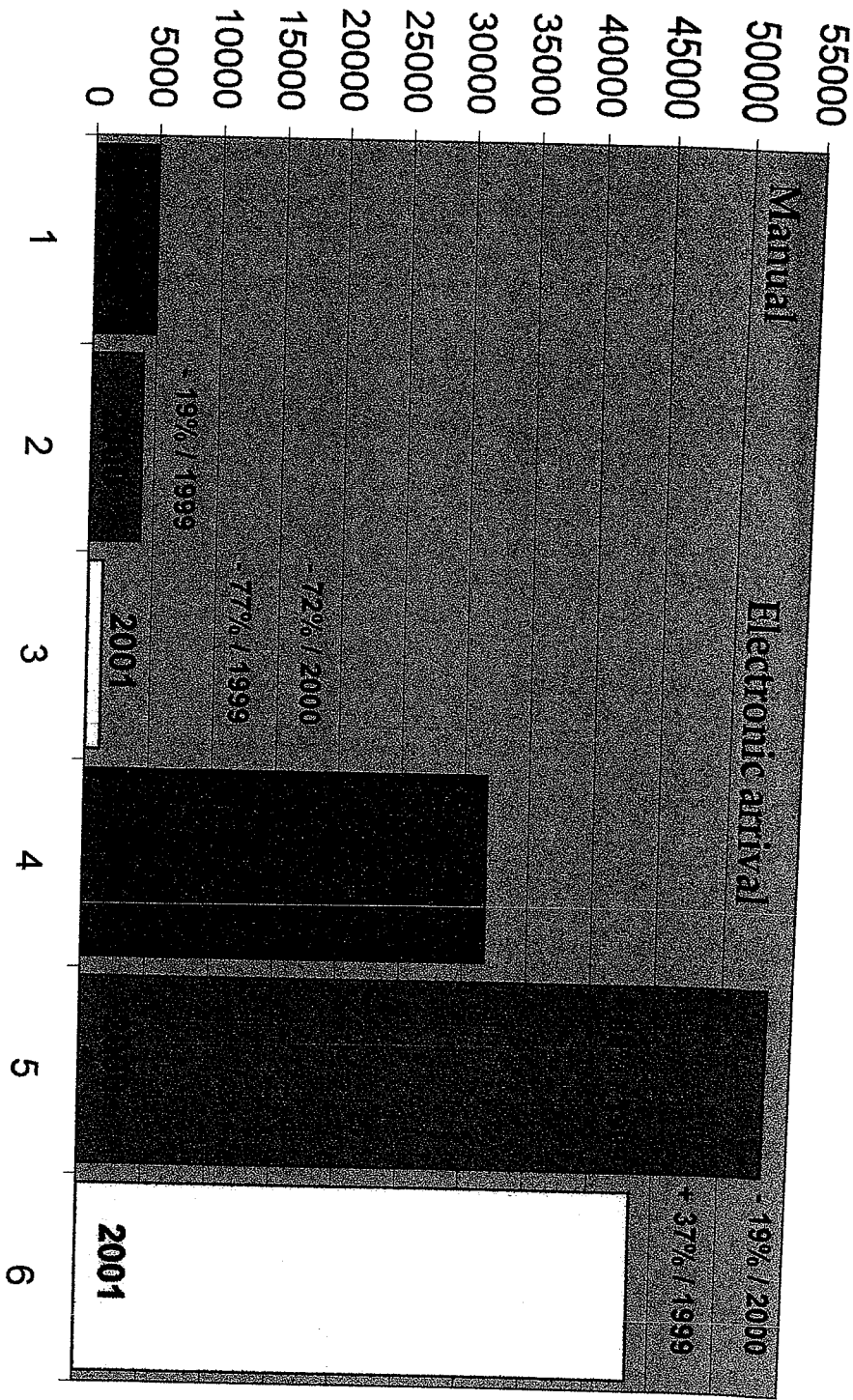
ANNEXE 10

STATISTIQUE DES IMPORTATIONS MANUELLES ET AUTOMATIQUES DE LITTÉRATURE GRISE AU CERN DES ANNEES 2000-2001

CURRENT AND RETROSPECTIVE PREPRINTS AND REPORTS in the CERN Library, 2001

2/Total of manual input and electronic arrival - Comparison 1999/2000/2001

4916 3987 1136 31699 53125 43280



ANNEXE 11

LISTE DES SOURCES EXPLOITEES AUX CERN EN 2001

**Grey literature
Sources**

13/ Comparison 1998/1999/2000/2001

	1998	1999	2000	2001
Importations to CERN (HEP)				
Shaky sources				
BLIC : British Library Inside Conferences (conf. articles)			2594	569
DISXA : Dissertations Abstracts / DataStar (theses)			309	0
Small sources incl. retrospective records				
DOE : US Dept. of Energy (reports)			1198	463
FERMI : FERMILAB, Batavia, IL (preprints)			2044	250
IN2P3 : France Institutes (preprints)			651	185
KEK : Tsukuba, Japan (preprints, reports, proceedings)			592	139
MATHDOC : Grenoble (theses, articles)			1227	530
MPARC : Texas, US, Austin Univ. (preprints)			2038	260
SPHT : Saclay, FR, CEA (preprints)				62
Small sources, current records only				
CERN EDS : CERN server (preprints, articles, theses, reports)	1163	3106	1658	794
DESY : Hamburg (preprints, theses)	86	481	1188	199
FIZ : Karlsruhe (conf. ann.)			853	713
ILLANT : Univ. Illinois (preprints)				229
ILLKT : Univ. Illinois (preprints)				386
JET : UK (preprints, reports)				478
JINR PREP : Dubna (preprints, reports)				196
KIMCONF : US (conf. ann.)			43	92
SLAC : Stanford, CA (conf. ann.)			405	470
TIPTOP : Bristol, IOP (conf. ann.)			114	214
WEIZMANN : Rehovot, Israel (preprints, articles)				22
Big sources incl. Retrospective records				
ING : INGENTA (publication references)				94
INSP : INSPEC (articles and publication references)	1035	1462	1319	1033
LANL EDS : LANL E-PRINT Los Alamos (preprints, articles, theses, reports)	24561	29715	32897	33308
LANLPUBL : LANL E-PRINT Los Alamos (publication references)			1467	1811
PROQUEST : US, North American Univ. (theses)			5466	3036
UNC : UNCOVER, Denver, Colorado Univ., publication references continued by INGENTA in 2001	8909	8451	979	251

Note : These tables don't include very small sources (Alice, Audio, CERN EDS-AGE, EPAC, ICHEP, INFN, ING, INTC)

	1998	1999	2000	2001
Exportations from CERN to SLAC				
CERN				
Increase	712	502	900	526
		-29%	79%	-42%