Cette étude présente une méthode de contrôle automatique des références des documents importés dans des bases de données. Conçu initialement pour la mise à jour de la base de littérature grise du Service de l'information scientifique du CERN, puis étendu à sa base de monographies, le programme *chkall* (« check all ») propose différentes applications pour contrôler les données de catalogage relatives aux documents (électroniques ou non) et pour en assurer la modification semi-automatique. Devant l'augmentation massive des données et la nécessité de les contrôler, ce programme, transposable à d'autres contextes, permet d'enrichir un catalogue et d'en augmenter la précision et la cohérence.

par CATHERINE CART
et INGRID GERETSLÄGER,
CERN, Genève

Méthode de contrôle automatique des données bibliographiques dans les bases de données du CERN

■ AU SEIN DE L'ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE (CERN), la section Gestion documentaire du Service de l'information scientifique (CERN-ETT-SIDM) a pour rôle de diffuser les travaux en physique des hautes énergies des chercheurs du monde entier.

La naissance du web, créé au CERN en 1989 par Tim Berners-Lee, puis l'arrivée en 1994¹ des serveurs (de publications, de prépublications, de thèses, etc.) ainsi que la nouvelle politique d'acquisition du Service de l'information scientifique (SIS)² ont eu pour conséquence une augmentation considérable des données électroniques à traiter³. Parallèlement, la recherche de la qualité optimale de celles-ci devint un souci majeur afin de satisfaire au mieux les besoins des chercheurs.

Ainsi, depuis quelques années, cette section utilise de plus en plus les nouvelles technologies pour accéder aux très nombreux documents électroniques diffusés par les serveurs et développer ainsi sa politique d'acquisition. Pour éviter d'avoir à procéder manuellement au catalogage des documents, le SIS s'est orienté peu à peu vers l'importation de données bibliographiques issues de laboratoires de tous horizons et provenant donc de bases de données ou de sources diverses accessibles par des sites web [1] [2] [3] [4].

La quantité croissante de ces importations (environ mille à mille deux cents documents sont importés chaque semaine dans notre base qui en contient aujourd'hui plus de quatre cent mille) a très vite soulevé le problème aigu de la correction et de la mise à jour des données de catalogage. Effectuer comme auparavant des modifications manuelles, notice par notice, champ par champ, ne pouvait plus être envisagé.

Le SIS fit alors appel aux outils proposés par le logiciel Aleph⁴ :

- **les filtres** [voir fig. 1] servant à vérifier des champs clés comme ceux qui génèrent des liens sur le web (champs relatifs aux références des publications par exemple). Mais cet outil se révéla insuffisant à cause du volume de notices à traiter alors qu'il ne peut fonctionner sur plus de 2.500 notices, et parce qu'il ne peut détecter certains types d'erreurs. Il restait donc un grand nombre de corrections manuelles à effectuer ;
- **les utilitaires** [voir fig. 2] permettant d'éditer massivement d'autres champs clés, choisis pour vérification de leur format. Ils sont apparus également inefficaces en raison de l'ampleur des corrections à faire manuellement et du temps nécessaire pour obtenir une base parfaitement fiable lors d'une recherche effectuée par les usagers sur l'interface web du catalogue ;
- **enfin des listes de références**, comportant des champs rejetés (hors format) avec leur correspondance avec des champs retenus (conformes)⁵, ont été créées pour formater automatiquement certains champs.

En complément de ces outils d'Aleph qui ne permettaient pas de faire fonctionner des listes de références et d'assurer une maintenance de tous les champs des notices, il s'imposait de concevoir un programme informatique spécifique⁶.

1 Le programme de contrôle automatique *chkall*

Ce programme a été conçu au départ pour mettre à jour la base de données de littérature grise du SIS, la base HEP (High Energy Physic). Il a été appliqué ultérieurement à la base des monographies. Il sert donc aujourd'hui à la mise à jour des deux bases.

Constitution d'un descriptif pour l'élaboration du programme

Chaque règle de catalogage dut être énoncée minutieusement pour chaque champ et sous-champ des notices selon la nature des documents : prépu-

blication, rapport, thèse, publication ou tout autre document relevant de la littérature grise.

Ce descriptif fut ensuite présenté au service CERN Document Server team, constitué d'informaticiens avec qui collabore le SIS⁷, qui choisit de traduire ce descriptif en langage PERL⁸ pour aboutir au programme définitif nommé *chkall*.

Fonctionnement du programme *chkall*

Appliqué à un ensemble plus ou moins grand de notices stockées dans un fichier, *chkall* détecte les erreurs de formatage de ces notices.

Il faut d'abord extraire de la base un fichier de notices à corriger, par le biais des filtres et des utilitaires [voir fig. 3]. Cette manière de procéder permet de choisir les données à corriger et de les extraire, totalement ou partiellement, soit par année ou par semaine, soit par type de document (prêtirage, publication, thèse, etc.), soit par type de champ seulement (on peut se limiter à des corrections ne portant que sur tel champ pour un ensemble de notices ou sur deux ou plus...).

Il faut ensuite lancer le programme *chkall* sur le fichier d'extraction. Cette opération permet non seulement le nettoyage hebdomadaire aisé des

Catherine Cart travaille à l'unité prêtirages au sein de la section Gestion documentaire du Service de l'information scientifique (SIS) du CERN, à Genève. Courriel <catherine.cart@cern.ch>

Ingrid Geretschlager est responsable de la section Gestion documentaire du Service de l'information scientifique du CERN. Courriel <ingrid.geretschlager@cern.ch>

CERN (Organisation européenne pour la recherche nucléaire), CH-1211 Genève 23, Suisse.

1 ArXiv.org e-Print archive / LANL, Los Alamos National Laboratory (Los Alamos, NM), avec plus de 170.000 prépublications et communications scientifiques produites depuis 1991, permet à tout scientifique de diffuser ses travaux électroniquement. La section CERN Document Server team a réalisé au CERN en 1994 un premier programme informatique d'importation spécifique de ces données. D'autre part, elle a créé par la suite un serveur propre au CERN (<http://cds.cern.ch/>).

2 Le CERN Document Server team a mis au point avec le SIS un deuxième programme informatique d'importation de données provenant de divers instituts, nommé *Uploader*. Ce travail a été présenté dans un article publié en 2001 dans cette revue [4].

3 De 1998 à 2001, le nombre de prépublications et de rapports importés par le service CERN-ETT-SI-DM est passé de 32.027 à 46.318.

4 Aleph 300, produit par la société Ex Libris, est un logiciel documentaire israélien. Système intégré modulaire, il propose des modules de catalogage, gestion des localisations, gestion des périodiques, maintenance et autorité, circulation, acquisition web serveur et client, opac, prêt entre bibliothèques.

5 Le fonctionnement de cet outil est développé en figure 7.

6 Parallèlement, d'autres outils ont été développés pour détecter des doublons, des erreurs dans les noms d'auteurs, etc.

7 Précisons toutefois que cet article n'a pas été écrit en collaboration avec ce service mais rédigé uniquement par des bibliothécaires.

8 Practical Extraction and Report Language. Ce langage de programmation, créé en 1986 par Larry Wall, peut interpréter des fichiers textes pour en extraire des données. Il est bien adapté aux procédures liées à la gestion de système.

-1- Exemple de filtres fonctionnant

S1 HEP 796 sw=200243	796 notices sont filtrées
S2 HEP 81 Limit S1 S ba 13	81 de ces notices figurent dans la base 13 (base des publications du CERN)
S3 HEP 5 Limit S2 R pr 0 99	5 de ces notices correspondent à des articles publiés dans un journal (champ PR) et comportent des mentions de pages
S4 HEP 60 Limit S2 R lkr PR PR	60 de ces notices correspondent à des travaux publiés dans un compte rendu de conférence (champ LKR) et comportent des mentions de pages
S5 HEP 65 s3 or s4	somme des notices de documents publiés
S6 HEP 16 s2 not s5	différence entre le nombre de notices figurant dans la base des publications et celui des notices de documents publiés

Ce type de filtres trouve dans la base des publications (ba 13) 16 notices avec une erreur, car sans pagination dans le champ de référence de publication (PR ou LKR).

Exemple du champ PR mal formaté :

\$ \$p Int. J. Mod. Phys. D11 (2002) 1171-1182

Correction manuelle à apporter au champ PR :

\$ \$p Int. J. Mod. Phys., D \$ \$v11 \$ \$y 2002 \$ \$c 1171-1182

Signification des codes :

\$ \$p : sous-champ relatif au titre du journal

\$ \$v : sous-champ relatif au volume

\$ \$y : sous-champ relatif à l'année de publication

\$ \$c : sous-champ relatif à la pagination

D'autres filtrages seront à faire pour détecter les autres erreurs portant sur d'autres champs ou sous-champs.

Exemple d'un filtre ne fonctionnant pas

S1 HEP 796 sw=200243	
S2 HEP 60 Limit S1 S lkr PR	recherche des sous-champs justes (PR précédés et suivis d'un espace)
S3 HEP 60 Limit S1 S lkr PR%	recherche des sous-champs faux (PR précédés mais non suivis d'un espace)

Ce filtre ne détecte pas les espaces. Cependant un espace en moins dans ce sous-champ aura une incidence sur le lien qui ne pourra se faire avec le compte rendu de la conférence sur le web.

-2- Exemple d'utilitaire (iutil 3/3) permettant l'extraction de tous les documents (1000 notices) importés en semaine 43 dans la base de littérature grise HEP

Utilities Library CERN

Write file of Document records

Create a file of document record information from a file of Document numbers.

Input file name : **sw43**

Retrieve following Document record fields :

ALL in the first field code will retrieve all the fields of the document record.

Field code : all	Field code :	Field code :
Field code :	Field code :	Field code :
Field code :	Field code :	Field code :
Field code :	Field code :	Field code :
Field code :	Field code :	Field code :

Sort input file by system number (Y/N) : y

Output file name : sw43all

File Format (A/E) : A

A=Aleph format, E=Edit format

Current/Previous record C

Hour job should start 8 Cancel operation

Cet outil permet l'obtention d'un fichier complet des notices demandées.

-3- Exemple de fichier extrait par un utilitaire

Ce fichier comporte une seule notice complète à corriger, présentant cinq erreurs notées en gras et italique et une information manquante

CER 2320421 BASE (base)	L 11
CER 2320421 LN (language)	L eng
CER 2320421 YR (année)	L 1994
CER 2320421 SW (semaine)	L \$ \$n \$ \$w200243
CER 2320421 ER (numéro électr.)	L hep-ph/9311372
CER 2320421 EM (adresse)	L Ed Stoeffhaas <ed@phenxe.physics.wisc.edu>
CER 2320421 TI (titre)	L Heavy Charged Higgs Signals at the LHC
CER 2320421 AU (auteur)	L Barger, V
CER 2320421 AU (auteur)	L Phillip, R J N
CER 2320421 AU (auteur)	L Roy, D P
CER 2320421 IM (édition)	L \$ \$n Wisconsin Univ. Dept. Phys. \$ \$d29 Nov 1993 \$ \$c9 p
CER 2320421 EXT (lien)	L \$ \$xhttp://documents.cern.ch/cgi-bin/setlink?
base=preprint&categ=hep-ph&id=9311372 \$ \$nAccess to fulltext document	
CER 2320421 RN (numéro de rapport)	L MADPH 790
CER 2320421 NI (note interne)	L LANL EDS
CER 2320421 PR (réf. Publ.)	L \$ \$pPhys. Lett., B \$ \$v324 \$ \$y1994 \$ \$c236-240

nouveaux documents quotidiennement importés, mais aussi le nettoyage rétrospectif de toute la base de littérature grise qui, au cours du temps, avait connu différentes règles de catalogage. En raison de l'impossibilité de mener à bien une telle tâche par une procédure manuelle, son harmonisation n'avait jamais pu être menée. Cela est d'autant plus important que notre base de données globale va prochainement adopter le format MARC 21⁹, le format actuel étant un format MARC personnalisé.

Résultat

Après avoir été appliqué au fichier d'extraction, le programme *chkall* édite trois fichiers de corrections à traiter différemment, comme le montre la figure 4 (extraits des trois fichiers obtenus à partir d'environ un millier de notices importées au cours de la semaine 43).

- Le fichier 1 comprend des messages précis d'erreurs portant sur tous les champs et sous-champs des notices. Les corrections sont à traiter manuellement à partir des numéros de notices donnés.

Exemple. En considérant le résultat obtenu à partir d'une seule notice, celle de la figure 3, on constate que le fichier 1 contient un seul message d'erreur à corriger manuellement :

E: SYSNO=2320421 PR {p,c} missing for base 13

Il indique qu'il manque un sous-champ dans le champ PR (référence de publication). En effet, le sous-champ \$\$p n'a pas été trouvé car il manque un \$. Cette erreur a des incidences importantes sur les liens hypertextes, développés en figure 6.

- Le fichier 2 comporte des champs corrigés à importer automatiquement dans la base.

Exemple. Le fichier 2, toujours obtenu à partir de la seule notice de la figure 3, contient des champs qui remplaceront automatiquement les champs faux, après versement de ce fichier sur Aleph sous Unix. Rien ne sera donc fait manuellement.

CER 2320421 AU L Phillip, R J N

CER 2320421 AU L Roy, D P

CER 2320421 BASE L 13

CER 2320421 IM L \$\$pMadison, WIS\$sn
Wisconsin Univ. Dept. Phys.
\$\$d29 Nov 1993\$\$c9 p

CER 2320421 RN L MAD-PH-790

Il affiche deux noms d'auteurs dans le but de les conserver (ils seront sinon écrasés par l'ajout du premier auteur (AU2) dans le fichier 3 ci-dessous). Il transforme la base 11 (base des prépublications) en base 13 (base des publications) puisqu'il s'agit là d'un article. Il met à jour le champ IM en le com-

plétant par la ville de l'éditeur et standardise le champ RN (numéro de rapport) par des tirets (conformément aux listes de références des champs RN, IM, évoquées en figure 7).

- Le fichier 3 comporte des ajouts d'informations ou de champs à importer automatiquement de la même manière.

Exemple. Le fichier 3 contient des informations manquantes (qui seront ajoutées automatiquement dans la base par soumission de ce fichier). Rien ne sera donc, là non plus, fait manuellement.

CER 2320421 AU2 L Barger, V

CER 2320421 SUL xx

L'auteur est déclaré en tant que premier auteur, donc placé dans un champ AU2, et la catégorie manquante (champ SU) est ajoutée (avec le contenu XX) afin d'être précisée ultérieurement. (AU2 sera présent en format MARC 21 dans les champs 100, entrée principale, et 700, fichier d'autorités auteurs).

2 Les difficultés rencontrées

Rédaction du descriptif et de sa traduction en PERL

Les difficultés relèvent principalement de la diversité des documents de littérature grise. De ce fait, la rédaction du descriptif et du programme PERL fut ardue et complexe. Par exemple, la mise au point d'un seul champ ou zone comme IM, relatif à l'édition (ville, éditeur, année d'édition) et à la collation, fut longue et délicate. Chaque éventualité pouvant se présenter dans ce champ devait être envisagée et mentionnée [fig. 5], chaque notion interdépendante des autres devant figurer dans un ordre pertinent pour que le programme fonctionne correctement.

Plusieurs étapes

Le programme doit être lancé plusieurs fois sur le même ensemble de notices. En effet, une première série de corrections peut engendrer de nouvelles modifications détectées par *chkall*. Ainsi, pour une notice relative à un document produit par le CERN, *chkall* mentionne dans le fichier 1 par exemple l'absence du champ DI (relatif à la division, propre aux documents CERN). Après l'ajout manuel de ce champ, le nouveau lancement de *chkall*, sur cette même notice, identifiera celle-ci comme appartenant au CERN et mentionnera, dans le fichier 2, le sous-champ spécifique manquant, SW=\$\$ya2002, servant à extraire la notice pour l'édition du rapport annuel du CERN.

Gestion des fichiers de corrections

Le traitement d'un champ répétitif est des plus délicats. Lors de l'extraction par les utilitaires ►

⁹ Format de catalogage pour l'échange des données, MARC 21 est le nom donné au résultat de la fusion, opérée en 1994, entre les formats US MARC et CAN/MARC. Il propose cinq formats : données bibliographiques, fichiers d'autorités, fonds, classification et renseignements communautaires. La bibliothèque du Congrès et la bibliothèque nationale du Canada en éditent les versions officielles anglaise et française.

-4- Extraits de fichiers hebdomadaires

Obtenus à partir d'un millier de notices importées au cours de la semaine 43

• **Extrait du fichier 1**
(fichier de messages d'erreurs contenant en moyenne 200 messages)

E: SYSNO=2346427 AF=collaboration needs to go to CO
E: SYSNO=2346610 AU2 is missing and neither CO,CA,TI1,CA1 exists
E: SYSNO=2346189 BASE=13 LKR exists but \$\$e missing!
E: SYSNO=2346285 BASE=13 PR exists but \$\$c missing!
E: SYSNO=2346326 LKR \$\$b or d missing
E: SYSNO=2346191 LKR \$\$e missing but BASE 13
E: SYSNO=2346836 NO1 does not exist
E: SYSNO=2346281 PR {p,c} missing for base 13
E: SYSNO=2346285 PR {p,c} missing for base 13
E: SYSNO=2346741 RN=CERN but DI is missing!
E: SYSNO=2346775 SW \$\$s not followed by n,h,r,m,p,d
E: SYSNO=2346741 SW \$\$s or \$\$w missing

• **Extrait du fichier 2**
(fichier de champs corrigés contenant en moyenne 2.000 champs)

CER 2346131 AU L delete
CER 2346192 CO L \$\$nDELPHI Collaboration
CER 2346483 IM L \$\$pArgonne, IL \$\$nArgonne Nat. Lab. \$\$d19
Oct 2002 \$\$cmult. p
CER 2346134 IM L \$\$psine loco \$\$nsine nomine \$\$d2001 \$\$c166 p
CER 2346134 NI L delete
CER 2346131 RN L CERN-SPSC-2002-033
CER 2346197 RN L VUB-TENA-02-06
CER 2336940 SW L \$\$sn \$\$w200243 \$\$ya2002
CER 2346788 TI L Acceleration of ULtra High Energy Cosmic
Rays : Cosmic Zevatrions?

• **Extrait du fichier 3**
(fichier de champs à ajouter contenant en moyenne 50 champs)

CER 2346836 IN L REPORT
CER 2346836 IN L THESIS
CER 2346131 NI L LANL EDS
CER 2346131 AU2 L Stachel, Johanna

**-5- Extrait du descriptif pour le champ IM relatif à l'édition
(ville, éditeur, année d'édition) et à la collation**

- si IM n'existe pas, ajouter ce champ obligatoire (pour BA=11, 14, 16, 19⁽¹⁾) dans le fichier 3 (fichier de champs à ajouter)
- si IM existe :
 - vérifier l'ordre suivant IM=\$\$p\$\$n\$\$d\$\$c⁽²⁾ :
 - si l'ordre est correct, ne rien faire
 - si l'ordre est incorrect, insérer ce champ dans le bon ordre, dans le fichier 2 (fichier de champs corrigés)
 - vérifier que IM soit complet, avec des doubles \$ et 4 sous-champs \$\$p\$\$n\$\$d\$\$c :
 - si IM est complet, ne rien faire
 - si IM est incomplet :
 - s'il manque un \$ ou si les lettres sont fausses, le signaler dans le fichier 1 (fichier de messages d'erreurs)
 - s'il y a le signe ø dans \$\$c, le remplacer par mult. p dans le fichier 2 (fichier de champs corrigés)
 - s'il manque \$\$d ou \$\$c, ajouter l'année de YR dans \$\$d (si ba=13) et mult. p dans \$\$c⁽³⁾ et verser ce champ corrigé dans le fichier 2 (fichier de champs corrigés)
 - s'il manque \$\$p ou \$\$n, pointer vers la liste de références pour compléter l'information :
 - si RN ne figure pas dans la liste de références mettre IM=\$\$psine loco⁽⁴⁾ \$\$nsine nomine \$\$ddate \$\$cpages (pour toutes les bases), puis mettre ce champ à verser dans le fichier 2 (fichier de champs corrigés)
 - si RN figure dans la liste de références, corriger en conséquence le champ IM sinon prendre le deuxième RN ou le troisième, etc., et mettre ce champ corrigé dans le fichier 2 (fichier de champs corrigés)
 - si RN n'existe pas, se référer au champ ER (numéro électronique)

(1) Base 11 : prépublications, base 14 : thèses, base 16 : documents/comités scientifiques, base 19 : rapports.

(2) \$\$p sous-champ relatif à la ville, \$\$n sous-champ relatif à l'éditeur, \$\$d sous-champ relatif à l'année d'édition, \$\$c sous-champ relatif à la collation.

(3) L'année de publication (IM\$\$d) doit être identique à l'année contenue dans le champ YR pour la base 13 (base des publications).

(4) La ville inconnue ne peut se traduire par s.l. tout comme l'éditeur inconnu par s.n., comme le voudrait la norme. En effet, il faut écrire les mots en entier pour une recherche pertinente possible. On trouverait sinon tous les mots contenant s ou l ou m dans le champ IM.

d'un tel champ à partir d'un fichier spécifique, il faut penser à cibler tous les champs multiples (NI=LANL EDS, NI=PROQUEST, etc.)¹⁰ et surtout ne pas se contenter du premier champ rencontré (NI=LANL EDS). Sinon, les champs justes d'origine seraient remplacés par les champs corrigés. De même, *chkall* réaffiche, dans les fichiers créés, les champs multiples justes, pour ne pas les perdre ; c'est pourquoi, dans l'exemple détaillé ci-dessus, les champs AU (auteurs) justes sont affichés dans le fichier 2, comportant les champs corrigés.

D'autre part, un champ non aligné, c'est-à-dire déplacé d'un espace dans un fichier à verser dans la base, peut provoquer de graves perturbations sur le catalogue. D'autres types d'erreurs peuvent bloquer la procédure d'importation d'autres fichiers.

Enfin, la gestion des fichiers 2 (fichiers de champs corrigés) et fichiers 3 (fichiers de champs à ajouter), à verser automatiquement dans la base, nécessite une attention particulière. En raison du nombre important de notices importées et traitées chaque jour, la capacité de traitement du système actuel nous impose un nombre limité de corrections à faire dans la base (problème de mise à jour des fichiers d'autorités). De ce fait, nous devons mettre en place et gérer des listes d'attente de fichiers de corrections de plus en plus nombreux en raison de l'accroissement des données importées. Or les notices sont modifiables dans la base de données par tout collègue qui peut être amené entre temps à ajouter ou à changer des informations¹¹. Aussi les fichiers de corrections seront versés plus tard sur des notices qui ne correspondent plus aux notices d'origine. Cela peut générer une importation d'erreurs (doubles champs, fausse base, etc.) qui nécessite un nouveau lancement de *chkall*.

Les atouts du programme *chkall*

Rapidité à détecter toutes les erreurs

L'application *chkall* a besoin d'environ une heure pour traiter un fichier portant sur les notices d'une année entière (environ 60.000 notices), six minutes pour le fichier des notices hebdomadaires et cinq secondes pour une dizaine de notices.

D'autre part, *chkall* traite tous les champs des documents du SIS, soit plus d'une vingtaine, ainsi que leurs sous-champs. Le type d'erreurs sur chaque champ et sous-champ est multiple. Sur le seul

champ IM (relatif à l'édition), *chkall* peut trouver plus d'une quinzaine de types d'erreurs ; par exemple, le nom d'une ville peut être mal orthographié ou inexact, la lettre du sous-champ peut être manquante ou fausse, l'emplacement de ce sous-champ peut être dans un ordre non conforme aux règles, etc. Il aurait bien sûr été impossible de détecter et de traiter manuellement ces erreurs à une échelle aussi importante.

Toutefois, bien que complexe dans son élaboration, le lancement seul de *chkall* est très simple. Il a permis un changement considérable dans la nature de notre travail, devenu ainsi plus spécifique et moins répétitif. Les modifications manuelles après relecture d'une notice ont complètement disparu.

Bonne répartition des tâches

Les fichiers 1 de messages d'erreurs peuvent être traités aisément par les nombreux stagiaires que nous accueillons. De cette façon, ces derniers se familiarisent très rapidement avec la base de données et avec les règles de catalogage. Les fichiers à soumettre (ceux des champs corrigés et des champs à ajouter) restent du domaine des initiés à Unix. En effet, une erreur dans leur manipulation peut engendrer des problèmes considérables sur la base de données, comme nous l'avons indiqué précédemment.

Souplesse du programme

Ce programme a été écrit en respectant l'ordre de chaque champ des notices. Ce qui rend possibles des ajouts ou des retrais en fonction de l'évolution des règles de catalogage ou des changements décidés. Il nécessite un suivi constant. Il n'est donc pas figé mais permet une mise à jour rapide et rétroactive.

D'autre part, ce programme peut être utilisé à la carte, pour tel type de champ par exemple, en réalisant des fichiers d'extraction limités à ce seul champ. Il est d'ailleurs possible, de cette manière, de nettoyer des bases complètement différentes de celles prévues à l'origine par le programme, comme celle, par exemple, des monographies. Trois moyens sont possibles pour y parvenir. On peut réaliser des fichiers d'extraction limités aux champs communs de la base HEP et de celle des monographies. On peut aussi réaliser des fichiers d'extraction de notices complètes mais retirer ensuite, dans les fichiers d'erreurs, les corrections qui ne concernent pas les monographies¹². Enfin, le programme peut être lancé en ajoutant à la commande une restriction pour certaines options non conformes à la base des monographies.

Valeur ajoutée par le SIS

Le programme *chkall* a permis d'assurer la qualité d'un éventail très large de documents importés, permettant aux lecteurs d'interroger l'interface web de manière optimale, rapide et pertinente. ►

¹⁰ NI : note interne indiquant la source de la notice. NI=LANL EDS indique que la notice provient (sous autorisation) de ArXiv.org e-Print archive de LANL, Los Alamos. NI=PROQUEST indique que la notice provient de US, North American Univ.

¹¹ Le catalogage est une fonction ouverte à tous au SIS.

-6- Création de liens hypertextes

Une erreur de \$ manquant dans le sous-champ du champ PR (relatif à la référence de publication), tout comme une erreur provenant d'une autre lettre que p,v,y,c dans les sous-champs, empêche la création d'un lien hypertexte sur l'interface web du catalogue du CERN (<http://weblib.cern.ch/>)

Le champ PR doit être structuré ainsi :

\$p titre du journal \$v volume \$y année de publication \$c pagination

Par exemple :

PR=\$p Phys. Rev., D \$v48 \$y 1993 \$c4779-4784

ne donnera pas sur le web le titre du journal ni le lien hypertexte vers l'article considéré.

Par contre, l'ajout du \$ manquant par le programme *chkall* donnera sur l'interface web du catalogue CERN le titre du journal et le lien hypertexte vers l'article recherché :

PR=\$p Phys. Rev., D \$v48 \$y 1993 \$c4779-4784

Quantum Mechanics, Common Sense and the Black Hole Information Paradox / Danielsson, U H ; Schiffer, M ;

gr-qc/9305012 ; 14 May 1993. - 15 p. - Abstract - Fulltext -

In: Phys. Rev., D : 48 (1993), pp.4779-4784 - Published version

Detailed record - Mark record

-7- Extrait de la liste de références des titres de journaux
(plus de 2.000 lignes de références), servant à formater le champ PR, relatif à la référence de publication

La colonne de gauche affiche les titres rejetés, non formatés, tels qu'ils peuvent arriver des serveurs dans le champ PR et la colonne de droite affiche les titres retenus. Donc, à partir d'un titre non formaté importé dans la base, *chkall* se réfère à cette liste et corrige le titre avec sa correspondance formatée en colonne de droite :

A & A	---Astron. Astrophys.
A & A	---Astron. Astrophys.
A A LETT	---Astron. Astrophys.
A A LETTERS	---Astron. Astrophys.
A AS	---Astron. Astrophys. Suppl. Series
A & A	---Astron. Astrophys.
A & A	---Astron. Astrophys.
A & AS	---Astron. Astrophys. Suppl. Series
AA	---Astron. Astrophys.
AAL	---Astron. Astrophys.
ABI TECHN	---ABI Techn.
ABITD7	---ABI Techn.
ACAD CIENC	---Acad. Cienc.
ACAD CIENC ZARAGOZA	---Acad. Cienc. Zaragoza
Etc.	

Exemple d'un champ PR avant sa correction par *chkall*:

PR=\$pA &A \$v387 \$y2002 \$c605

Le même champ PR après sa correction par *chkall*:

PR=\$pAstron. Astrophys. \$v387 \$y2002 \$c605

Extrait de la liste de références des villes et éditeurs

(plus de 1800 lignes de références), servant à formater le champ IM

À partir de l'existence d'un numéro de rapport (champ RN), *chkall* corrige le champ IM en lui donnant sa correspondance (ville et éditeur) qu'il trouve dans la colonne de droite. Il peut aussi bien ajouter IM s'il n'existait pas, le corriger s'il était faux ou le compléter s'il était incomplet :

AAEC-TM	---\$pLucas Grenoble Heights \$nAustr. Nucl. Sci. Techn. Organ.
AC-INT	---\$pAachen \$nAachen TU 3. Inst. Phys.
ACT	---\$pHouston, TX \$nHouston Univ. Adv. Res. Cent. The Woodlands
AD-RHIC	---\$pUpton, IL \$nBrookhaven Nat. Lab.
ADP	---\$pAdelaide \$nAdelaide Univ. Dept. Phys. Math. Etc.

Exemple d'un champ IM (avec le champ RN=AD-RHIC présent dans cette même notice) avant sa correction par *chkall*: IM=\$pBrokaven \$n lab. \$d2000 \$cmult. p

Le même champ IM après sa correction par *chkall*: IM=\$pUpton, IL \$nBrookhaven Nat. Lab. \$d2000 \$cmult. p

Extrait de la liste de références des numéros de rapports,
servant à formater le champ RN

chkall corrige les numéros des documents en les standardisant selon la colonne de droite :

LA-UR	---LAUR
LPTENS	---LPT-ENS
LUNFD6	---LUNFD-6
MADPH	---MAD-PH
MD-PP	---UMD-PP
MD-PT	---UMD-PT
Etc.	
MD-TR	---UMD-TR
MIS	---MPI-MIS
MP-ARC-00	---MP-ARC-2000
MPIMIS	---MPI-MIS
MSUHEP	---MSU-HEP
OXFORD-NUCL-PREP	---OUNP
Etc.	

Exemple d'un champ RN avant sa correction par *chkall*: RN=MD-PT-85-001

Le même champ RN après sa correction par *chkall*: RN=UMD-PT-85-001

Les corrections et ajouts d'informations dans les notices enrichissent les données et permettent de faciliter l'accès à l'information sur le catalogue.

Une référence de publication formatée par *chkall* selon la norme ISO 4 (norme d'abréviation de mots de titres) permettra la préparation d'un lien dynamique sur l'interface web¹³. Le lecteur pourra ainsi accéder directement à l'article publié dans le journal, si la bibliothèque est abonnée à ce journal électronique [voir fig. 6].

Le programme *chkall* est rattaché d'ailleurs à des listes de références [voir fig. 7], comme évoqué en début d'article, pour formater et standardiser automatiquement les champs correspondant par exemple aux titres des journaux, aux villes et éditeurs ou encore aux numéros de rapports. À partir d'une information non conforme aux fichiers d'autorités, le programme renvoie aux listes de références et permet d'obtenir la forme retenue des informations spécifiques à ces listes.

4 Un programme utilisable dans d'autres services de documentation

Il a fallu deux années de travail pour parvenir à la mise au point complète de ce programme. Elle s'est traduite par des élaborations très précises, des tests, des transformations constantes et de nombreuses mises à niveau dans le catalogue.

Cet outil a répondu aux objectifs fixés et les a même dépassés. En effet, il permet d'une part la mise à jour des notices importées quotidiennement, de plus en plus nombreuses. Il a de plus rendu possibles l'harmonisation et l'uniformisation rétrospective de tout le catalogue de la littérature grise ainsi que de la base des monographies. Ce qui permettra ainsi la conversion cohérente des données au format MARC 21, prévue prochainement, et la constitution du nouveau catalogue¹⁴. De ce fait nous poursuivons non seulement notre travail d'équipe pour la maintenance de ce programme mais nous lui apportons déjà quelques modifications pour qu'il puisse fonctionner dans sa nouvelle version sous MARC 21. Vu le succès qu'il a remporté dans notre service, les autres sections du SIS commencent à prendre modèle sur cette technique performante (pourtant déjà expérimentée auparavant en d'autres circonstances [5] [6]) et à mettre en place elles aussi de tels programmes.

Dans sa nouvelle version, *chkall* pourrait aussi être facilement utilisable et ajustable par d'autres services de documentation utilisant le format MARC 21.

NOVEMBRE 2002, Texte mis à jour en juillet 2003

Références

- [1] **Automatisation partielle du traitement de la littérature grise dans le service d'information scientifique du CERN** / Catherine Deroche, Ingrid Geretschläger (dir.). Genève : CERN, 1998. 59 p. DESS : Sci. Inf. : ENSSIB/Univ. Lyon 1 <<http://preprints.cern.ch/archive/electronic/cern/preprints/thesis/thesis-98-019.ps.gz>>
- [2] **Automatisation du traitement des documents CERN** / Catherine Cart, Ingrid Geretschläger. Genève : CERN, 1999. 6 p. <<http://preprints.cern.ch/archive/electronic/cern/preprints/open/open-99-068.pdf>>
- [3] **Traitement de publications CERN de l'intranet : importation automatique/semi-automatique de publications d'expériences CERN dans le catalogue de la bibliothèque** / Philippe Ricanet, Jocelyne Milan (dir.), Ingrid Geretschläger (dir.). Genève : CERN, 1999. 75 p. Maîtrise : Documentation : Univ. Lyon 3 <<http://documents.cern.ch/archive/electronic/cern/preprints/thesis/thesis-99-064.pdf>>
- [4] **Le traitement informatisé de ressources électroniques au Service de l'Information Scientifique du CERN** / Nathalie Pignard, Ingrid Geretschläger, Jocelyne Jerdelet. *Documentaliste - Sciences de l'information*, 2001, vol. 38, n° 1, p. 24-34.
Version anglaise : High Energy Phys. Libr. *Webzine*, 3 (2001) 3
Également accessible sur Internet en versions française : <<http://documents.cern.ch/cgi-bin/setlink?base=other&categ=sis&id=sis-2000-531>> et anglaise : <<http://documents.cern.ch/cgi-bin/setlink?base=other&categ=sis&id=sis-2001-092>>
- [5] **Quality control in online databases** / Edward T. O'Neill, Diane Vizine-Goetz. *Annual Review of Information Science and Technology* (ARIST), 1988, n° 23
- [6] **A Pascal interface for data validation of CDS/ISIS database records** / M. K. Herala, A. S. Ravi, T. B. Rajashekar. *Information Studies*, Oct. 1996, 2(4), p. 263-272

12 Comme par exemple celles touchant les informations de conférences ou comptes rendus, ainsi que celles touchant la collation pour les monographies disponibles sur Internet.

13 Une fois structuré, le champ de référence d'une publication est utilisé par un programme (adopté par l'American Physical Society pour ses journaux électroniques) pour calculer l'algorithme de l'URL de la page de l'article. Cela évite d'avoir à ajouter des DOI abstraits à la notice.

14 <http://cdsweb.cern.ch>