Automatisation du traitement des documents CERN

[Article rédigé en français – This paper is written in French]

Catherine Cart*- Ingrid Geretschläger*

**Laboratoire Européen de Physique des Particules (CERN)*
1211 Genève 23
Suisse
catherine.cart@cern.ch
ingrid.geretschlager@cern.ch

RESUME

Dans le cadre de l'établissement de la Liste des publications du CERN, la bibliothèque a pour objectif de substituer, à l'actuel dépouillement manuel des périodiques et comptes rendus de conférences, des procédures automatiques. Celles-ci reposent sur l'importation de références de publication ou de notices bibliographiques et sur l'interrogation de bases de données.

MOTS-CLEFS : document électronique, automatisation, traitement manuel

ABSTRACT

Within the context of producing the List of CERN Publications, the library is aiming to replace the manual analysis of periodicals and conference proceedings by automatic procedures. These rely on importing references of publication or bibliographic notices and by querying databases.

KEYWORDS: electronic document, automation, manual input

1. Etat des lieux

Dans un centre de recherche fondamentale comme le CERN, les résultats des expériences sont consignés dans des pré-tirages, à raison d'environ 2000 documents par an pour la seule production CERN.

Pour permettre à l'ensemble de la communauté scientifique d'accéder aux pré-tirages dans les meilleurs délais, ils sont intégrés au plus vite dans notre base de données (logiciel documentaire Aleph), consultable sur le Web, http://alice.cern.ch.

Ces pré-tirages sont destinés à être publiés, le plus souvent dans des périodiques ou dans des comptes rendus de conférences. Ces articles, avec d'autres documents, font ensuite l'objet d'une publication annuelle dans la Liste des publications du CERN (ci-après <<la Liste>>) qui reflète la production intellectuelle du Laboratoire. Elle revêt une importance décisive pour la Direction du CERN, les délégués des Etats membres et des pays observateurs et sert à l'élaboration de la politique budgétaire.

1.1 *Documents électroniques*

Aujourd'hui, ces pré-tirages proviennent pour moitié de sources électroniques. C'est pourquoi nous les appelons documents électroniques. Ils sont transférés dans la base de données de la bibliothèque par le biais de deux serveurs : celui de la bibliothèque (créé en 1994) et celui de Los Alamos. Ce dernier, également créé en 1994, est utilisé par la majorité des physiciens du monde entier pour soumettre leurs documents. L'exportation de ses données est gratuite. C'est pourquoi la bibliothèque utilise ce serveur pour compléter sa base de données.

Pour l'établissement de la Liste, les références de publication sont ajoutées manuellement aux notices bibliographiques. Elles sont obtenues par le travail de dépouillement manuel des périodiques et compte rendus de conférences. Les auteurs, en effet, ne donnent pas ces informations à la bibliothèque lors de la publication de leurs documents.

1.2 <<*By-passed documents*>>

Outre ces pré-tirages, il en existe d'autres que la bibliothèque ne reçoit jamais. Nous les appelons <<*by-passed documents*>> (documents hors-circuit).

Ils sont localisés au cours du travail de dépouillement. Leurs notices bibliographiques, y compris leurs références de publication, sont ensuite entièrement saisies manuellement.

2. Statistiques

La Liste 1997 contient 1872 documents (83 % d'articles et 17 % de livres, comptes rendus, thèses, *proposals*[1]). Elle se compose de 1570 articles (57 % de documents électroniques et 43 % de <<*by-passed documents*>>.

Les divisions de recherche produisent la majorité des articles (1281 articles sur un total de 1570). Cependant, elles ne nous communiquent que 56 % de leurs articles, les 44 % restants représentent les <<*by-passed documents*>>

Concernant les divisions des accélérateurs, sur un total de 289 documents**,** 61 % des articles sont des documents électroniques et 39 % sont des <<*by-passed documents*>>.

3. Résultats actuels

Le traitement entièrement manuel appliqué aux <<*by-passed documents*>> pourrait être évité si les auteurs soumettaient leurs pré-tirages au serveur de la bibliothèque ou à celui de Los Alamos. De plus, le document électronique, dès son entrée dans la base de données, offre sur le Web son contenu intégral, à la différence des <<*by-passed documents*>>. En omettant de soumettre son pré-tirage par voie électronique, l'auteur prive ainsi tous les lecteurs du contenu du document.

Concernant les documents électroniques, il suffirait que l'auteur informe la bibliothèque de la publication de son pré-tirage pour limiter le travail de dépouillement manuel. Celui-ci n'obéit d'ailleurs à aucune démarche méthodique, car il concerne uniquement les publications acquises par la bibliothèque. Il reste donc aléatoire.

Par ailleurs, deux personnes à temps complet sont occupées à l'établissement de la Liste. Cet investissement en temps n'est pas négligeable dans une période de restrictions des ressources budgétaires et humaines.

Enfin, ce travail est d'autant plus lourd que le nombre de pré-tirages est en constante augmentation (il a doublé depuis 1994).

Pour conclure, outre le traitement des <<*by-passed docum*ents>> entièrement manuel et très fastidieux, celui des documents électroniques, partiellement manuel, reste encore beaucoup trop lourd.

Dans ces conditions, il est clair que le dépouillement doit progressivement céder le pas à des procédures automatiques.

[1]Procès-verbaux des expériences édités par les comités scientifiques du CERN

4. Mise en place de procédures automatiques

4.1 *En amont : le catalogage*

Par le passé, au niveau du catalogage, de nombreux champs étaient «inventés» et les sous-champs trop souvent ignorés. Curieusement, le champ Collaboration était choisi, alors qu'AACR2 propose le champ Collectivité. Il en est de même pour le sous-titre, signalé lors de la saisie par « espace : espace» et non pas par un sous-champ (séparant le titre propre du sous-titre). On imagine aisément les problèmes de comparaison de fichiers et d'importation de notices qui en découlent.

De plus, les bases de données utilisées pour différencier la nature du pré-tirage sont trop nombreuses (base 11 : pré-tirage soumis à un périodique ; base 12 : pré-tirage présenté à une conférence ; base 13 : pré-tirage publié dans un périodique ou dans un compte rendu de conférence) .

Enfin, d'ici l'an 2000, la bibliothèque adoptera le format USMARC pour mettre en place la nouvelle version du logiciel documentaire, Aleph 500. Des réunions sont organisées afin d'harmoniser et de normaliser nos règles de catalogage.

La diminution de la part de catalogage passe aussi par une collaboration plus étroite avec les auteurs et leurs secrétariats. Ceux-ci sont régulièrement invités à des réunions d'information sur les procédures de soumission des pré-tirages au cours desquelles l'accent est mis sur la grande qualité d'information qui en résulte. Enfin, l'existence d'un document officiel, intitulé <<Circulaire administrative 29 : Principes et procédures régissant les publications et rapports du CERN et les autres publications résultant de travaux effectués au CERN>>, énonçant les règles CERN en matière d'information scientifique auxquelles les auteurs devraient se soumettre, est régulièrement rappelée dans le bulletin d'information hebdomadaire du CERN.

4.2 *En aval*

Pour automatiser le traitement des pré-tirages (documents électroniques et <<*by-passed documents*>>), la bibliothèque fait appel surtout aux ressources existantes, telles que les bases de données primaire et secondaires.

4.2.1 *La base de données primaire SLAC : procédure opérationnelle*

La bibliothèque échange régulièrement ses pré-tirages CERN, présents dans sa base de données, avec le *Stanford Linear Accelerator Center (SLAC)*, aux Etats-Unis. Elle les lui transfère électroniquement avec leurs numéros de série uniques. D'autre part, SLAC dépouille les articles des périodiques pour mettre à jour les références de publication de ses pré-tirages comme des nôtres. De ce fait, jusqu'au début de 1997, la bibliothèque recherchait manuellement les référence de ses documents électroniques dans la base de données SLAC, http://www-spires.slac.stanford.edu et les ajoutait manuellement.

En septembre 1997, un programme d'importation de ces références a été mis au point par notre ingénieur informaticien. Ce programme identifie les documents par interrogation du

numéro de série, puis importe, après correction, leurs références de publication dans la base de données CERN. Plus de 500 références ont été ainsi importées pour la Liste 1997.

4.2.2 *Les bases de données secondaires UNCOVER et INSPEC : projets en cours*

Diverses bibliothèques américaines dépouillent un certain nombre de périodiques scientifiques et regroupent les documents dépouillés dans la base de données UNCOVER, sur Web et Telnet.

Nous en avons choisi une dizaine, les plus courants édités par *l'American Physical Society* . Nous avons fait appel à un ingénieur informaticien de l'université de Sienne pour étudier sur la base de ces titres un programme de comparaison entre UNCOVER et le CERN portant sur des mots significatifs du titre des documents électroniques CERN (en tenant compte des particularités de LaTeX)[2]. En fonction du résultat, la référence de publication serait importée automatiquement dans la base de données CERN.

Cette procédure automatique, gratuite, nous permettrait aussi d'importer les références de publication des documents électroniques non CERN (soumis aux périodiques) que la bibliothèque possède dans sa base de données.

Un autre projet en cours est celui de l'interrogation de la base de données INSPEC[3]. 66 % des périodiques que possède la bibliothèque du CERN sont dépouillés par IEE ainsi que les articles des comptes rendus de conférences.

Il suffirait d'importer d'INSPEC les références de publication des documents électroniques CERN et la notice bibliographique entière des <<*by-passed documents*>>. Avec l'aide d'une étudiante en DESS informatique documentaire de l'université de Lyon I, l'élaboration d'un prototype de programme de comparaison est à l'étude. En raison de la complexité de ce programme, les recherches ne pourraient se faire que par accès Telnet DATASTAR/DIALOG, payant[4], au moyen d'une équation de recherche.

Les notices bibliographiques localisées par cette équation seraient formatées selon les règles de catalogage de la bibliothèque puis examinées avant d'être importées ou rejetées.

5. Conclusion

La transition du traitement manuel vers le traitement automatique des documents électroniques de la Lise a commencé avec la procédure SLAC et reste en projet avec celles d'UNCOVER et d'INSPEC. Les <<*by-passed documents*>> seront peut-être importés dans l'avenir par la procédure INSPEC.

---

[2]Formateur de texte utilisé pour la rédaction de formules mathématiques

[3]INSPEC est produit depuis 1969 par IEE, *Institution of Electrical Engineers*, Londres. INSPEC regroupe *Computer and Control Abstracts, Physics Abstracts* et *Electrical and Electronics Abstracts* entre autres et offre un accès payant Web et Telnet via DATASTAR/DIALOG. La base de données contient aujourd'hui 5,8 millions de notices et affiche une croissance annuelle de 250.000 notices. Sa mise à jour est hebdomadaire. 82 % des articles proviennent des journaux.

[4]Coût estimé pour 1998 : 6000 US $

Les statistiques nous ont permis de situer tout d'abord le terme «document électronique» dans son contexte. Elles en montrent clairement les limites. Elles nous ont aussi conduits à nous interroger sur les procédures appliquées auparavant et à les repenser. Leur renforcement par des programmes informatiques devrait permettre une automatisation accrue du dépouillement. Une automatisation à 70 % serait considérée comme un succès en termes de coût - efficacité et de valorisation du travail. L'avenir nous confirmera ou non le bien-fondé de notre stratégie. Elle nous amènera inévitablement à relever encore d'autres défis qui ne rendront notre travail que plus dynamique.

6. Bibliographie

[CER94] CERN. GENEVE. DIVISION DU PERSONNEL, *Circulaire administrative 29 : Principes et procédures régissant les publications et rapports du CERN et les autres publications résultant de travaux effectués au CERN*, CERN, 1994.

[GOO98] GOOSSENS, M., LE MEUR, J. Y., *Afficher les documents scientifiques sur le Web*, Cahiers GUTenberg, v. 28-29, p181-196, 1998.

[LEM98] LE MEUR, J. Y., BRUGNOLO, F., *The personal virtual library*, 7th International World Wide Web Conference, Brisbane, Australia, 14-18 Apr. 1998. – Rapport CERN-OPEN-98-019.