Le traitement informatisé de ressources électroniques au Service de l'Information Scientifique du CERN

Nathalie PIGNARD, doctorante au GRESEC[1], actuellement stagiaire au CERN
GRESEC, Université Stendhal, Institut de la Communication et des Médias, av. du 8 mai 1945, 38130 Échirolles
nathalie.pignard@u-grenoble3.fr

Dr. Ingrid GERETSCHLÄGER, responsable de la section Gestion documentaire
CERN[2], Service de l'Information Scientifique, section Gestion Documentaire, 1211 Genève 23, Suisse
ingrid.geretschlager@cern.ch

Jocelyne JERDELET, responsable de l'unité Prétirages
CERN, Service de l'Information Scientifique, section Gestion Documentaire, 1211 Genève 23, Suisse
jocelyne.jerdelet@cern.ch

Résumé : Nous présentons une méthode automatique d'importation de données mise en œuvre au Service de l'Information Scientifique, SIS, du CERN. Ce programme informatique, baptisé *Uploader*, permet d'importer dans le catalogue de la bibliothèque du CERN des notices bibliographiques et le texte intégral de documents provenant de diverses sources sur Internet. Ces bases de données concernent la littérature grise en physique et dans les disciplines voisines (par exemple DOE, KEK, Math-Doc, TipTop, etc.). Cette politique d'acquisition, qui met en avant le traitement informatisé des ressources électroniques, soulève quelques réflexions sur l'augmentation du nombre de documents collectés et sur l'élargissement des domaines traités. Le souci constant d'enrichir ces données et d'en faciliter l'accès aux utilisateurs, sur un mode hypertextuel, conduit à une évolution des métiers de la gestion documentaire.

Mots-Clés : Littérature grise - Traitement informatique - Importation documentaire - Ressources électroniques - Politique d'acquisition

Soumis à Documentaliste - Sciences de l'information

Novembre 2000

[1] Groupe de Recherche sur les Enjeux de la Communication, Université Stendhal, Grenoble III
[2] Organisation Européenne pour la Recherche Nucléaire

Contexte : du papier à l'électronique

Depuis plus de quarante ans, le Service de l'Information Scientifique (SIS) du CERN[3] [1] collabore avec des instituts de recherche et des universités[4] du monde entier à la diffusion des travaux effectués par les scientifiques.
Ainsi, la bibliothèque du CERN reçoit régulièrement, via des listes de diffusion, les documents écrits par les chercheurs de ces laboratoires et universités. Les documents, sous forme papier, sont ensuite numérisés afin de fournir aux utilisateurs un accès, sur le Web, à ces travaux de recherche.
Aujourd'hui, cette pratique tend à s'affaiblir ou tout au moins à se transformer. En effet, la littérature grise en sciences, et plus particulièrement en physique, circule de plus en plus sous forme électronique. Après avoir assumé pendant plusieurs années la diffusion des documents sur les deux supports, certains instituts font aujourd'hui le choix de l'électronique qui présente des avantages indéniables par rapport au papier : économie des coûts, diffusion facilitée, consultation du texte intégral des documents à distance, possibilité d'enrichir le catalogue et l'accès en ligne aux documents à moindre coût, etc. Maurice B. Line [2] souligne encore d'autres attraits au document électronique : "*Les principaux critères d'efficacité sont la rapidité de la fourniture du document, la fiabilité (la probabilité d'obtenir un document à partir de la ou des sources approchantes) et la facilité d'utilisation.*"
A l'ère de la "bibliothèque virtuelle", les documents papier se font donc de plus en plus rares et les auteurs préfèrent généralement soumettre directement leurs travaux sous forme électronique.
En outre, la plupart des laboratoires ont accès aujourd'hui au Web et beaucoup ont cessé la diffusion de leurs documents papier (Fermilab[5] aux Etats-Unis, Nordita au Danemark, …) et invitent les bibliothèques scientifiques et les chercheurs à consulter leurs pages Web ou leurs bases de données.

Face à cette évolution, les politiques d'acquisition doivent être reconsidérées et s'adapter aux nouvelles normes de circulation de l'information scientifique [3]. Dans cette optique, le Service de l'Information Scientifique, SIS, du CERN, et plus particulièrement la section *Document Management* (Gestion Documentaire), oriente progressivement ses choix vers le traitement informatisé des ressources électroniques. Depuis quelques années, des travaux d'étude et de recherche sont régulièrement engagés sur ce sujet au sein du Service de l'Information Scientifique du CERN [4], [5], [6], [7], [8].
Dans ce contexte nouveau, le problème qui se pose est celui de la consultation multiple des bases de données : pour trouver un document, un chercheur devra nécessairement consulter plusieurs sources, ce qui est un travail long et fastidieux. Pour faciliter la recherche et offrir aux scientifiques une seule interface d'interrogation et de visualisation, le SIS a choisi de rapatrier le maximum de documents électroniques dans sa base de données [9].

[3] Organisation Européenne pour la Recherche Nucléaire
[4] Par exemple GANIL (Grand Accélérateur National des Ions Lourds, Caen), DESY (Deutsches Elektronen Synchrotron, Hambourg), LAPP (Laboratoire d'Annecy-le-Vieux de la Physique des Particules, Annecy), MPI (Max Planck Institut, Garching), GSI (Geschellschaft für Schwerionenforschung, Darmstadt), RAL (Rutherford Appleton Laboratory, Chilton), DAPNIA (Département d'Astrophysique, de Physique des Particules, de Physique Nucléaire et de l'Instrumentation Associée, Saclay), SFB (Sonderforschungsbereich, Technische Univ. Berlin), Budker Institut for Nuclear Physics (Novosibirsk), Meisei Univ. (Tokyo), etc.
[5] Les sigles et abréviations utilisés dans ce document sont expliqués à la fin de l'article

Ainsi, le support informatique de la bibliothèque du CERN a mis au point, il y a une année, un programme baptisé *Uploader*, permettant d'importer de façon automatique des notices bibliographiques provenant de diverses sources [10].
L'intérêt de cet outil est triple puisqu'il devait permettre :
- de pallier à la diminution des listes de diffusion que les instituts de recherche se font de plus en plus rares à envoyer sous forme papier, en recueillant directement sur leurs sites les travaux des chercheurs,
- d'élargir le nombre de documents obtenus précédemment sous forme papier dans ces différents laboratoires et universités,
- mais aussi d'explorer de nouvelles bases de données proposant des documents très intéressants pour les physiciens du CERN et d'enrichir ainsi le catalogue de la bibliothèque.

L'importation automatique de notices électroniques

Le fonctionnement de l'*Uploader*

A partir d'un fichier de données provenant de n'importe quelle base de données ou page Web, le programme *Uploader* formate les notices pour les adapter au catalogage de la bibliothèque du CERN [Annexe 1].
Pour chaque nouvelle source importée, des fichiers de configuration sont créés afin de permettre la mise en forme des champs des notices d'origine au format MARC[6] utilisé par la base du CERN[7].
Ce programme propose également d'autres fonctionnalités comme la mise à jour de notices existantes, la recherche dans le catalogue du CERN pour repérer les notices déjà présentes avant l'importation, etc.

Le choix des sources

Le choix des catalogues à explorer s'est effectué selon plusieurs critères. Le premier a consisté à consulter les sites Web de tous les instituts pour lesquels le CERN reçoit encore les travaux des chercheurs sous forme papier et de voir si ces sites proposaient en ligne ces mêmes documents.
Cette analyse a révélé que la quasi totalité des instituts avaient fait le pas de la mise en ligne de leurs documents, mais de façon plus ou moins approfondie. Cette étude a également montré que le CERN ne recevait en moyenne, via les listes de diffusion, qu'un tiers des travaux produits dans ces instituts. Deux explications peuvent être avancées : il est probable que pour des raisons de coût, les laboratoires effectuent une sélection des documents avant de les envoyer aux instituts partenaires; de plus, les listes de diffusion ne sont pas toujours tenues à jour et le CERN reçoit donc de moins en moins de documents par ce biais.
L'utilité d'importer de façon automatique les documents depuis les sites Web de ces laboratoires était donc indéniable, mais s'est trouvé confrontée à d'autres problèmes, d'ordre technique, que nous commenterons plus loin.

---

[6] Machine Readable Cataloguing

[7] Chaque "configuration" est caractérisée par trois fichiers principaux. Deux fichiers permettent de définir la structure des champs de la notice d'origine, en vue de leur extraction. Un troisième fichier sert à créer la nouvelle notice, avec les champs adéquats : à partir des données d'origine, différentes commandes sont mises en oeuvre pour adapter ces informations aux exigences du formatage utilisé dans le catalogue du CERN.

D'autres sources ont été envisagées dans des domaines où la base de données de la bibliothèque est encore peu développée. C'est le cas notamment de disciplines comme les mathématiques (avec Math-Doc à Grenoble ou mp_arc à Austin), ou encore pour des types de documents comme les thèses (avec par exemple Proquest[8], base hébergée par Data Star).

Deux méthodes pour traiter les données localisées sur Internet

Les différentes sources observées appartiennent à deux grandes catégories : les pages Web et les bases de données en ligne. Leur mode de fonctionnement est totalement différent et leur traitement via l'*Uploader* diffère sensiblement.

Les pages Web des instituts de recherche

Les laboratoires et instituts de taille moyenne, qui ne disposent pas d'une base de données en ligne, proposent généralement sur leur site des pages Web présentant les travaux de leurs chercheurs (le plus souvent les thèses). Les fonctionnalités offertes sont très sommaires puisque ces sites ne proposent pas de possibilité de recherche. Généralement, les notices sont triées par type de document (thèses, preprints, etc.) et parfois par année. Avec ce système, leur nombre reste souvent limité. C'est pourquoi il n'a pas paru intéressant, dans la plupart des cas, de créer une configuration pour chacune de ces sources, une soumission manuelle des notices et du texte intégral étant tout aussi rapide. Le second argument, que nous développerons plus loin, est que le manque de stabilité des pages Web est un obstacle à la mise en place de configurations pour l'importation automatique des documents.
La question principale concerne le suivi de ces pages : comment être averti d'une modification, voire de l'arrivée d'une nouvelle notice ? Les services d'alerte sur ces sites sont encore rares et seules deux des sources analysées proposent ce service : TipTop[9] (I.O.P, à Bristol) pour les annonces de conférences et mp_arc pour les preprints en mathématiques. Une autre solution consiste à poser une veille sur ces pages et d'être ainsi informé de leur évolution. Environ quatre-vingt veilles ont été posées pour couvrir la production d'une trentaine d'instituts [Annexe 2].

Les bases de données

Les bases de données en ligne offrent souvent la possibilité de mener des recherches multicritères. Mais à l'inverse des pages Web décrites ci-dessus, il est généralement impossible de poser une veille sur la page des résultats générée par la recherche. Il est donc difficile d'importer de façon très régulière (toutes les semaines par exemple) les notices nouvellement entrées dans ces bases, sauf pour celles qui proposent un service d'alerte.
La méthode définie pour l'importation depuis ces bases consiste donc à faire des recherches annuelles pour obtenir l'ensemble des notices de l'année écoulée. L'inconvénient de cette recherche annuelle est que l'on obtient des notices bibliographiques avec plusieurs mois de retard.

[8] *Proquest Digital Disserations* est une version gratuite, mais limitée, de *Dissertation Abstracts International* (UMI). Elle contient des notices de thèses soutenues dans les universités d'Amérique du Nord et dans 200 autres du monde entier. La période couverte comprend l'année en cours et l'année précédente.
[9] TipTop, a *Unified Physics Resource* est le résultat d'une initiative privée entre TipTop (Kenneth Holmlund, Mikko Karttunen et Günther Nowotny) et la base de données PhysicsWeb / IOP(Institute of Physics Publishing, Bristol). TipTop est maintenu depuis 1998 par I.O.P et s'adresse à la communauté de recherche en physique.

Un autre obstacle fréquemment rencontré lors de l'étude de ces bases de données concerne la mise en forme de la page de résultats. La plupart du temps, les résultats s'affichent sous forme de liste; pour davantage d'informations sur l'une des entrées, l'utilisateur doit cliquer sur le lien hypertexte qui le mène à la notice détaillée. Fréquemment, la page de résultats ne fournit pas suffisamment d'informations sur chacune des entrées. Il est courant par exemple que seul le premier auteur d'un document soit donné (ex. DOE, *Department of Energy*) [Annexe 3.1] ou que le titre soit tronqué (ex. base de données CITHER [11]) [Annexe 3.2]. Dans ces cas-là, une importation peut se révéler difficile voir impossible.
En outre, le catalogage étant spécifique à chaque sorte de document, avec des différences plus ou moins marquées, plusieurs configurations sont parfois nécessaires pour une même base de données proposant divers types de documents.
Ainsi, de juillet à octobre 2000, 14 configurations ont été créées pour 9 bases de données.

Les problèmes rencontrés

L'instabilité des pages Web

Les pages sur le Web sont marquées par une certaine instabilité qui revêt plusieurs formes.

Instabilité dans le temps tout d'abord, puisque les pages peuvent à tout moment disparaître, ce qui pose notamment problème lorsque l'on importe, en plus de la notice, le lien vers le texte intégral des documents stockés sur le site du laboratoire concerné.

Instabilité des pages dans leur structure, ensuite. En effet, pour plusieurs configurations, les balises html présentes dans le fichier source des pages Web permettent de délimiter facilement les champs constitutifs d'une notice. Cependant, ces balises ne sont pas toujours régulières et stables d'une page à l'autre, voire sur une même page. En effet, la plupart du temps, les pages se présentent sous forme de texte libre, et les champs n'ont pas toujours de structure commune (espaces, tabulateurs, interlignes, etc.). Les contraintes codées imposées par les bases de données sont donc inexistantes; or la mise en page libre ne se prête guère à l'élaboration d'une configuration.

Instabilité dans la présentation des notices et des champs, enfin. La raison de cette instabilité est certainement que les pages de ces laboratoires de recherche sont souvent créées et mises à jour par des secrétariats ou des personnes non professionnelles de la documentation, ce qui provoque une hétérogénéité dans les champs, plus fréquente - et plus gênante - pour le champ des auteurs (dans la base mp_arc, Austin, par exemple) [Annexe 4].
Certaines bases offrent même la possibilité à des personnes extérieures de saisir de nouvelles notices (ex. TipTop pour les annonces de conférences ou encore Los Alamos[10] qui laisse aux auteurs le soin de soumettre leurs travaux), ce qui provoque de grandes irrégularités et un manque d'homogénéité dans la présentation des documents.

[10] ArXiv.org e-Print archive / LANL, Los Alamos National Laboratory (Los Alamos, NM) contient depuis 1991 plus de 170000 prétirages et communications scientifiques en physique, mathématiques et informatique avant leur publication et offre le texte intégral des documents.

Le travail manuel de vérification reste nécessaire

Cette instabilité dans les pages Web est peu compatible avec la structure très rigide exigée par le catalogage de bibliothèque. C'est pourquoi, l'un des soucis premiers du Service de l'Information Scientifique étant d'offrir aux utilisateurs une base propre et homogène, tout le travail manuel de vérification et de validation des notices importées est conservé.

Il est indéniable que l'utilisation de ce programme offre un gain de temps considérable par rapport aux soumissions manuelles et élargit le nombre de documents rendus disponibles (voir les statistiques pour l'année 2000 en Annexes 6).

Toutefois, ces procédures nécessitent de consacrer du travail à la mise en place des configurations, à la sélection des bases de données, à la recherche des notices présentant un intérêt pour notre catalogue et enfin à l'importation, la validation et la correction de ces données avant leur intégration dans le catalogue. La mise en oeuvre de ce procédé d'importation est donc essentiellement intéressante pour des bases de données importantes.
De plus, l'instabilité des pages Web évoquée plus haut impose un suivi régulier des sources et la mise à jour des fichiers de configuration. Nous pouvons donc en conclure qu'avec ce type d'outil, le travail des bibliothécaires persiste mais change, et s'oriente davantage vers la correction des notices importées que vers la création manuelle de nouvelles entrées.

Cette évolution dans l'activité du SIS du CERN se traduit également par le souci constant d'offrir de la valeur ajoutée, c'est-à-dire l'ajout d'informations (dans les notices) ou de services aux utilisateurs (via l'interface Web du catalogue) rendant ainsi les données plus riches et l'accès à l'information facilité.

La valeur ajoutée par le SIS du CERN

La "valeur ajoutée" fournie par la bibliothèque concerne aussi bien des corrections sur les données importées que des mises à jour de certains champs ou encore la mise en place de liens hypertextuels entre différentes informations [Annexe 5].

Liens entre les notices de la base

Les contributions à une conférence sont, sur la version Web de la base de données du CERN, liées entre elles ainsi qu'au compte-rendu de la conférence. Ainsi, d'un clic, un utilisateur peut, à partir de la notice d'un article, avoir accès aux informations concernant la conférence, à son compte-rendu, ou encore aux autres papiers soumis au même événement. Ce champ est dynamique dans la mesure où une modification faite dans la seule notice du compte-rendu s'affiche sur le Web pour toutes les contributions soumises à cette conférence.
De même, la notice d'un article publié peut être liée à celle d'un reprint et/ou d'un compte-rendu et/ou d'une revue. Plusieurs renvois sont possibles pour un seul article, mais leur gestion est délicate, car la moindre erreur empêche le lien de fonctionner.

Uniformisation et standardisation

Un travail important concerne l'uniformisation et la standardisation de champs comme celui des auteurs par l'adoption d'une translittération, notamment pour les noms russes ou nordiques[11]. L'objectif est de faciliter la recherche d'un nom aux utilisateurs en uniformisant leur orthographe.

La standardisation concerne également les références de publication : l'abréviation des noms de revues répond à des normes (ISO 4). Grâce à un fichier de correspondances, les différentes formes orthographiques connues des revues sont automatiquement modifiées pour afficher la forme désirée. L'objectif de cette standardisation est d'assurer, pour la version Web de la base, le lien vers ces revues électroniques lorsque la bibliothèque en possède la licence.

Ajout d'informations

Certaines bases de données ne proposent qu'un nombre limité d'auteurs (par exemple une trentaine pour les preprints en provenance de la base de Los Alamos). Le SIS se charge donc d'ajouter, lorsque c'est le cas, les auteurs non mentionnés dans la notice en extrayant les données à partir du ficher texte attaché à la notice bibliographique. Ceci est notamment très intéressant pour les grandes expériences qui affichent souvent plus de 500 auteurs dans les prétirages.

Pour certaines notices, des informations non contenues dans les données d'origine, mais facilement déductibles sont ajoutées. C'est le cas par exemple pour des documents concernant des expériences du CERN : des champs comme l'affiliation, la division et le nom de l'accélérateur liés à cette expérience sont créés.

Quelle légitimité pour ce type de procédé ?

Jusqu'à aujourd'hui, l'expérimentation de cette nouvelle forme d'acquisition des ressources électroniques ne s'est déroulée que sous forme de tests. Il convenait en effet dans un premier temps de s'assurer que ces importations étaient techniquement possibles et surtout qu'elles présentaient un intérêt réel en terme de gain de temps et d'enrichissement du catalogue.
Dorénavant se pose la question de la légitimité de ce type de procédure. En effet, ces importations ne peuvent se dérouler dans l'ombre, sans en avertir les laboratoires directement concernés. C'est pourquoi, si la période de tests se révèle positive, le SIS a décidé d'informer officiellement tous ces instituts et de leur demander l'autorisation d'importer une partie de leurs notices bibliographiques, ainsi que le texte intégral des documents - lorsqu'ils le proposent - dans la base du CERN.
Cette démarche a déjà été engagée pour certains instituts l'année dernière et s'est révélée encourageante : des universités comme celle de Cornell, des laboratoires comme Fermilab ou des bases de données comme Inspec[12] et FIZ ont donné leur accord pour que la bibliothèque du CERN procède à des importations automatiques depuis leurs catalogues.

---

[11] Par exemple, les terminaisons russes -ii, -ij, -y sont uniformisées en -y; les formes en ö, oe, o, Ø sont systématiquement orthographiées Ø, etc.

[12] Inspec, banque de données bibliographiques produite par l'*Institution of Electrical Engineers*, contient presque 7 millions de notices depuis 1969. Cette base dépouille la plupart des revues et comptes-rendus des conférences anglophones en sciences.

En contrepartie, le SIS du CERN propose d'insérer dans les notices un lien vers le site d'où sont extraites les données, afin que l'importation soit transparente et non dissimulée aux utilisateurs. Ce lien renvoie vers la page d'accueil de la base ou bien vers le texte intégral du document s'il est disponible.

Conclusion

Cet outil informatique permettant l'importation de ressources électroniques répond à la volonté du SIS d'offrir aux utilisateurs un catalogue le plus exhaustif possible dans les domaines de recherche traités au CERN. Ainsi, l'utilisateur sait qu'il peut trouver dans la base de données non seulement les papiers des travaux effectués au CERN, mais également ceux d'instituts comme Dapnia, KEK, SLAC, etc. qui mènent des recherches complémentaires.
En outre, le souci est d'offrir une base de données "propre", d'où le travail de vérification et de modification des notices importées, la quantité ne devant pas se substituer à la qualité, ce qui peut être un risque dans ce type de politique d'acquisition.
La "valeur ajoutée" fournit par les bibliothécaires du SIS du CERN est donc primordiale et cette nouvelle forme d'acquisition des données ne se résume pas à de simples rapatriements de données à l'aide d'un programme informatique.

Aujourd'hui, plus de 90% des notices créées dans le catalogue[13] du CERN le sont de façon électronique. Parmi elles, seul 3% représente des soumissions sur le serveur du CERN par des chercheurs ou leurs secrétariats, le reste étant le fruit d'importations automatiques, telles que décrites dans cet article [Annexe 6].

Sur un plan plus général, cette nouvelle forme de politique d'acquisition adoptée par le SIS du CERN depuis quelques années est un moyen de pallier au manque de mise en application des discours utopiques sur la constitution d'un catalogue collectif en littérature grise. En effet, depuis une trentaine d'années, l'idée de créer une base de données commune incluant tous les catalogues des grandes bibliothèques est régulièrement discutée. Aujourd'hui, l'un de ces projets se développe activement; il s'agit de l'*Open Archives Initiative*, auquel le SIS du CERN va sans doute prendre part [Annexe 7].
Malheureusement, ces projets se trouvent la plupart du temps confrontés à divers obstacles dès leur lancement : des problèmes techniques (il est nécessaire d'adopter des normes communes, etc.), auxquels s'ajoutent souvent la lenteur et le manque de volonté politique à concrétiser ces propositions. C'est pourquoi, aujourd'hui, certaines bibliothèques scientifiques comme celle du CERN, sont amenées à se constituer, par leurs propres moyens, un catalogue suffisamment étoffé pour satisfaire leurs utilisateurs.

[13] Ces statistiques concernent la base de littérature grise qui comprend les preprints, les articles, les rapports et les thèses.

ANNEXES

Annexe 1 : retour texte

Un exemple d'importation : la base de données de l'institut *KEK*

* La notice d'origine (obtenue par la base de données KISS - KEK Information Service System)

199827167 KEK Preprint 98-167
Ohuchi, N.; Tsuchiya, K.; Ogitsu, T.; Ajima, Y.; Qiu, M.; Yamamoto, A.; Shintomi, T.(KEK, Tsukuba)
Magnetic field measurements of a 1-m long model quadrupole magnet for the LHC interaction region
[Scanned images][The first page]

* La notice telle que formatée pour la base du CERN

eng
1998
$$k 199827167
Magnetic Field Measurements Of A 1-m Long Model Quadrupole Magnet For The Lhc Interaction Region
Ohuchi, N
Tsuchiya, K
Ogitsu, T
Ajima, Y
Qiu, M
Yamamoto, A
Shintomi, T
$$n KEK $$p Tsukuba $$d Oct 1998 $$c mult. p
$$x http://www-lib.kek.jp/cgi-bin/img_index?199827167 $$n Fulltext
KEK-Preprint-98-167

* La notice telle que présentée via l'interface Web du catalogue du CERN

Magnetic Field Measurements Of A 1-m Long Model Quadrupole Magnet For The Lhc Interaction Region / Ohuchi, N; Tsuchiva, K; Ogitsu, T; Ajima, Y; Qiu, M; Yamamoto, A; Shintomi, T;
KEK-Preprint-98-167. - Tsukuba : KEK , Oct 1998. - mult. p. - Fulltext -
Detailed record - Mark record

Annexe 2 : retour texte

Les services d'alerte et la pose de veilles

Les services d'alerte

Certains sites analysés proposent des services d'alerte. Leur fonctionnement est simple : régulièrement (généralement chaque semaine) les nouvelles notices entrées dans la base sont envoyées, par email, aux personnes qui se sont inscrites à ce service. I.O.P offre cette possibilité sur son site *Physics Web* pour les annonces de conférences. De même, les *Mathematical Physics Archives* (mp_arc), gérées par l'université d'Austin (Texas) propose ce service.

Cette forme de liste de diffusion peut également se combiner à un autre service proposé par la base de données : il s'agit du dépôt d'un "profil" sous la forme d'une équation de recherche (par mots clefs, type de documents, périodes). L'équation de recherche s'effectue en principe de façon automatique tous les jours et les résultats sont envoyés par courrier électronique à chaque fois que des notices correspondant au profil sont ajoutées dans la base. Le courrier électronique affiche les notices de façon structurée avec un lien vers leur fichier texte (il est généralement possible de définir le format des notices recherchées). Nous les traitons à l'aide d'une configuration dans l'Uploader. Ce type de service est notamment utilisé par le SIS pour des bases de données comme FIZ ou Inspec.

La pose de veilles

Pour les pages n'offrant pas de service d'alerte tel que décrit ci-dessus (à savoir pour la majorité des sites observés), des veilles ont été posées sur toutes les pages jugées intéressantes et susceptibles d'évoluer. Par "veille", nous entendons l'observation automatique de pages Web par un robot. Nous avons choisi de poser ces veilles via le site Web Mind-It[14] qui propose ce service gratuitement. Cet outil parcourt régulièrement les adresses URL (uniform ressource locator) et détecte tout changement et intervention sur ces pages : modification (ajout, correction, suppression de données), migration d'adresse, disparition de la page. Tout changement est signalé par une icône; de plus, les modifications intervenues sur la page sont mises en surbrillance, ce qui permet de les repérer rapidement. Grâce à Mind-It, il est possible de créer plusieurs veilles, de les regrouper dans des dossiers, de leur donner des titres, et surtout de définir la fréquence de l'alerte, et son mode (alerte par email, ou à l'aide d'une icône sur la page même du site, etc.).
Cette forme de veille nécessite le lancement régulier (par exemple chaque mois) de MindIt. Nous soumettons ensuite toutes les nouvelles notices une à une au serveur CERN EDS. Cette procédure demande autant de travail qu'une saisie directe dans la base, mais a l'avantage de transférer et de stocker le fichier du texte dans ce serveur. Le serveur étant stable, le fichier est accessible et conservé.
Ainsi, à défaut de pouvoir créer un système d'importation automatique depuis ces sites, la pose d'une veille permet de suivre l'évolution de ces pages et de l'apparition de nouveaux documents.

[14] MindIt / NetMind, http://mindit.netmind.com/

Exemple de veilles posées via le site Mind-It

Annexe 3 : retour texte

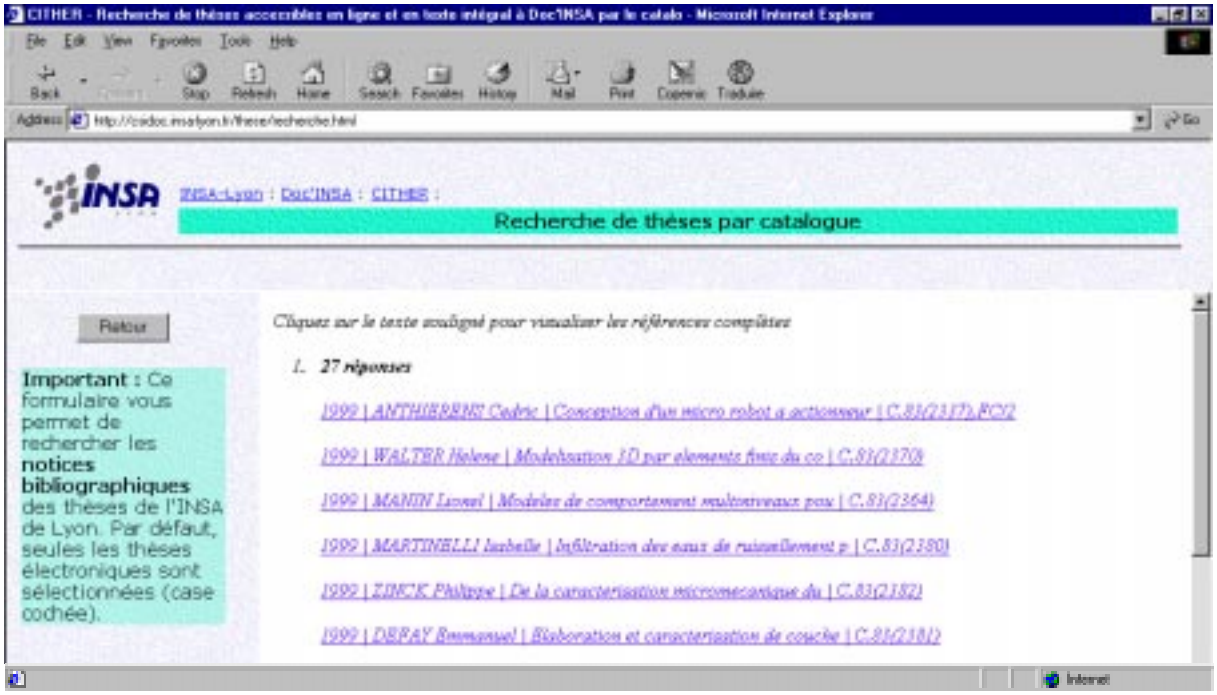
Exemples de problèmes rencontrés dans les sources pour la mise en place d'une configuration

Annexe 3.1 - Exemple du résultat d'une recherche dans la base de *DOE*

Seul le premier auteur est indiqué; pour obtenir les autres auteurs et plus de détails sur cette entrée, il faut cliquer sur le lien hypertexte du document.

Annexe 3.2 - Exemple de résultat d'une recherche dans la base *CITHER* retour texte

Le titre est tronqué (pour l'obtenir en entier, il faut cliquer sur le lien hypertexte de la notice). L'importation de ces notices est impossible.



Annexe 4 :

retour texte

Plusieurs champs "auteurs" extraits de notices provenant de mp_arc

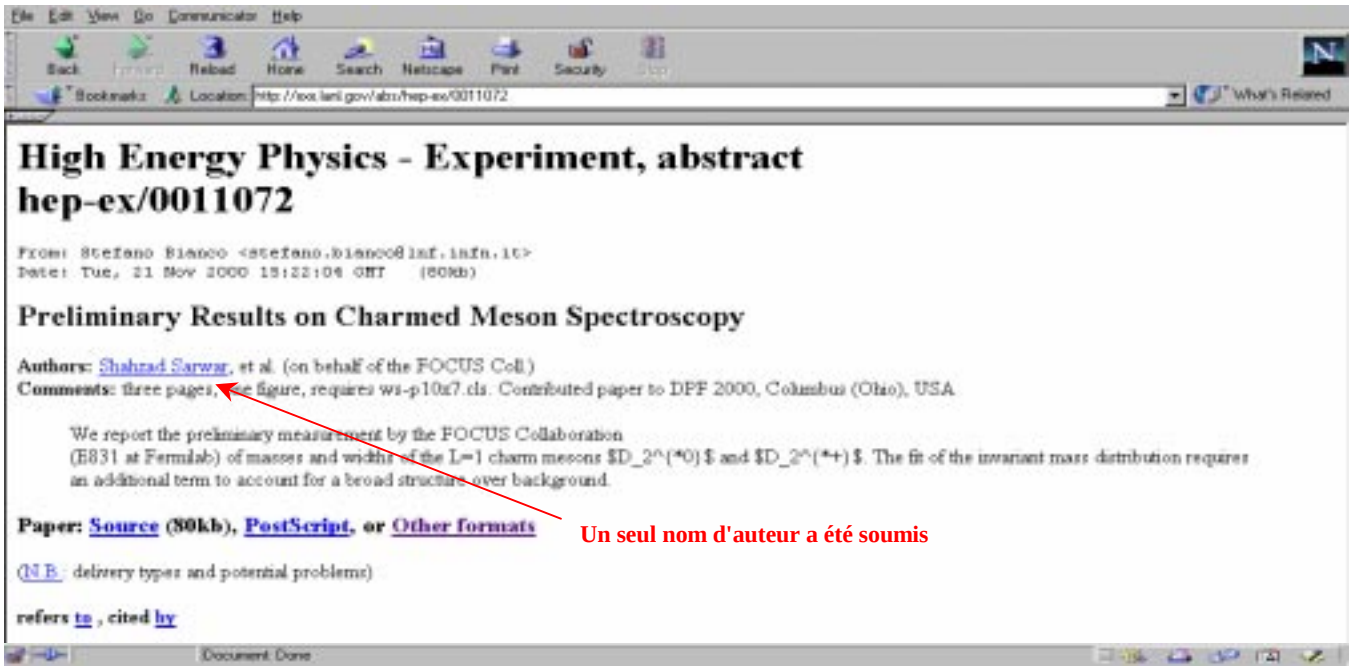
Champs des auteurs des notices entrées dans la base mp_arc pour la semaine du 19 au 26 octobre 2000. Nous pouvons observer de grandes irrégularités générant des difficultés à créer une configuration efficace pour leur importation. Une vérification et éventuellement une correction manuelle de chaque notice importée est donc obligatoire.

- Pavel Exner, Alain Joye
- A. Jorba
- J.Bricmont, A.Kupiainen, R.Lefevere
- Tai-Peng Tsai and Horng-Tzer Yau
- Werner Fischer, Hajo Leschke, Peter Mueller
- Bleher P., Ruiz J., Schonmann R.H., Shlosman S., Zagrebnov V.

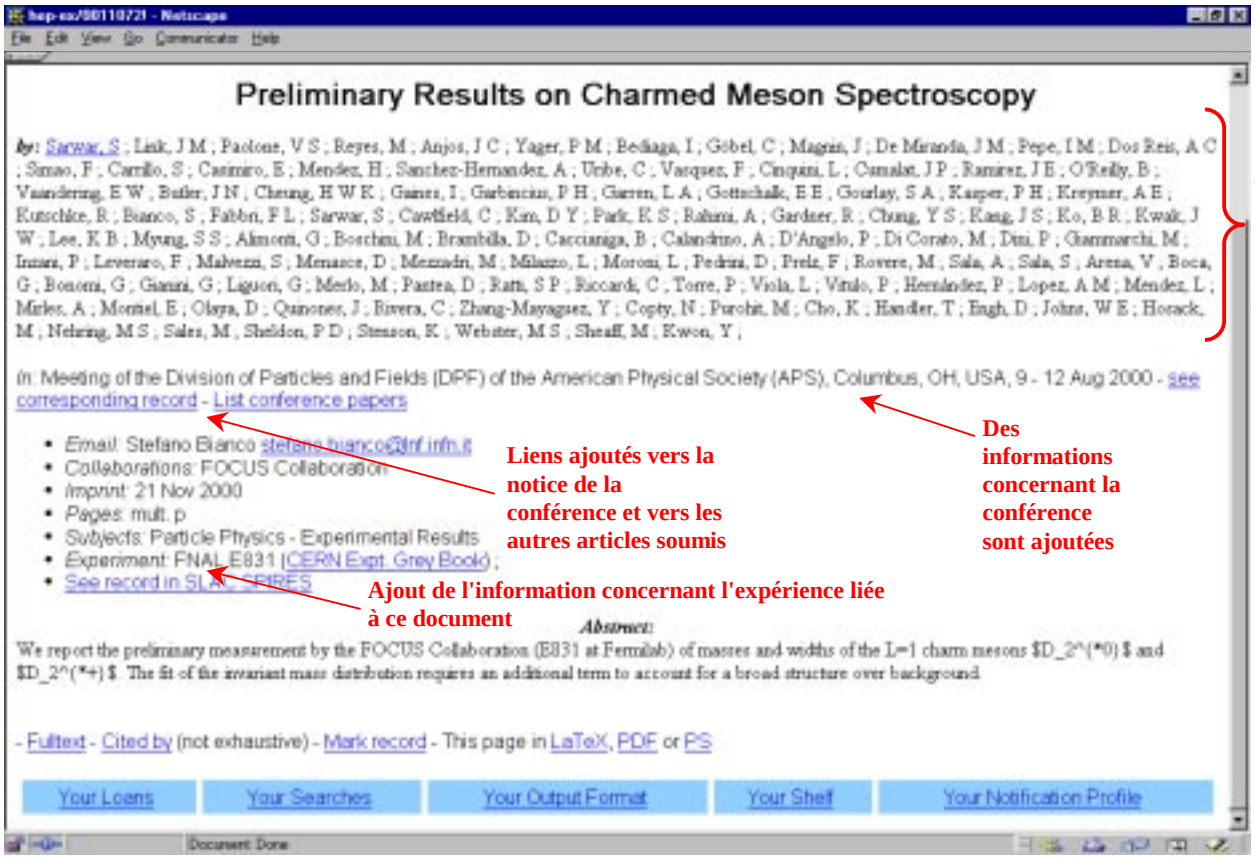
Annexe 5 : retour texte

La valeur ajoutée : exemple de l'importation d'une notice depuis le serveur de preprints de Los Alamos

Preprint soumis par les auteurs au serveur de LANL Los Alamos



La même notice dans le catalogue CERN SIS avec les valeurs ajoutées



Annexe 6 : retour texte

Statistiques : pourcentage de notices entrées manuellement et électroniquement dans la base du CERN de janvier à novembre 2000

Base de données de littérature grise : articles, preprints, thèses, rapports

Nombre total de notices entrées dans la base de données de littérature grise de janvier à novembre 2000 = **env. 53000**

Collecte de documents	**Sources**	**Nombre de notices**	**Pourcentage**
Saisies manuelles	Documents papier et listes	**4300**	**env. 8%**
Importations automatiques électroniques	Serveur CERN (soumission par le SIS, les auteurs des documents et leurs secrétariats)	1500	env. 3%
	Los Alamos	29000	env. 55%
	Autres (INSPEC, SLAC, etc.)	4200	env. 8%
	A titre de tests	14000	env. 26%
	Total des importations	**48700**	**env. 92%**
Total des notices entrées dans la base		**53000**	**100%**

Note : la base de données du SIS du CERN contient plus de 350000 notices

Annexe 7 :

retour texte

Le projet *Open Archives Initiative* (http://www.openarchives.org)

Le projet *Open Archives Intiative* répond à un appel lancé en juillet 1999 par Paul Ginsparg (initiateur de la base de preprints *e-Print archive* à Los Alamos), Rick Luce (LANL, Bibliothèque) et Herbert Van de Sompel (LANL, Bibliothèque). Leur volonté est de mobiliser un groupe de chercheurs et bibliothécaires européens dans l'optique d'établir un service universel destiné à l'auto-archivage de publications scientifiques par leurs auteurs.

L'*Open Archives Intiative* a déjà donné lieu à des rencontres et des réalisations concrètes : la réunion de Santa Fe (Nouveau Mexique) les 21 et 22 octobre 1999, qui a donné naissance à la « convention de Santa Fe », le rassemblement du 3 juin 2000 à San Antonio au Texas et celui des 18-20 septembre 2000 à Lisbonne. La prochaine rencontre de l'*OAI* aura lieu du 22 au 24 mars 2001 au CERN [12].

La convention de Santa Fe [13] a établi un certain nombre de principes de base, et notamment des recommandations précises pour l'implémentation d'interfaces permettant la récupération de métadonnées de chaque archive.

De plus, un site a été créé, et un logiciel permettant l'auto-archivage de manière interopérable a été développé par le département d'informatique de l'université de Southampton en Angleterre.
L'objectif de l'*OAI* à terme est que de multiples bibliothèques, par l'adoption de normes communes et d'un modèle de notice minimale, ouvrent l'accès de leurs catalogues et s'échangent leurs données respectives sans modifications locales lourdes.

Sigles et abréviations

Instituts et laboratoires de recherche

DOE	U.S. Department of Energy, Washington, DC
Fermilab	Fermi National Accelerator Laboratory, Batavia, IL
KEK	High Energy Accelerator Research Organization, Tsukuba, Japon
Nordita	Nordisk Institut for Teoretisk Fysik, Danemark
SLAC	Stanford Linear Accelerator, Stanford, CA

Base de données ou projets en cours

CITHER	Consultation en Texte Integral des Thèses en Reseau, INSA de Lyon
FIZ	Fachinformationszentrum Physik, Karlsruhe
Inspec	Information Service in Physics, Electrotechnology and Control
Math-Doc	Cellule de Coordination Documentaire Nationale pour les Mathématiques, Univ. Grenoble 1
mp_arc	Mathematical Physics Archives, Texas univ., Austin, TX

Références

[1] http://library.cern.ch

[2] Accéder ou acquérir, une véritable alternative pour les bibliothèques ? / Maurice B. Line *In* : BBF, 1996, t. 41, n° 1

[3] Quelle politique documentaire pour l'acquisition de liens Internet en bibliothèque ? / Isabelle Bontemps, Bernard Calenge (dir.). Lyon : ENSSIB, 1999. 67 p. Mémoire d'étude : D.C.B.
http://www.enssib.fr/bibliotheque/documents/dcb/bontemps.pdf

[3] Le traitement de la littérature grise à la bibliothèque du CERN / Isabelle Collignon, Ingrid Geretschläger (dir.). Genève : CERN, 1998. DEUG-DIST : I.U.P./Univ. Lyon 1

[4] 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. D.E.S.S. Sci. Inf. : ENSSIB/Univ. Lyon 1
http://preprints.cern.ch/archive/electronic/cern/preprints/thesis/thesis-98-019.ps.gz

[5] Automatisation du traitement des documents CERN / Catherine Cart, Ingrid Geretschläger. 1999. - 6 p. *Soumis à :* Document Numérique
http://preprints.cern.ch/archive/electronic/cern/preprints/open/open-99-068.pdf

[6] 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

[7] Comparative and statistical analysis between the CERN conference database and three other bases / Nathalie Pignard, Ingrid Geretschläger (dir.), Jocelyne Jerdelet (dir.). Genève : CERN, 1999. 53 p. Maîtrise Information Communication : Univ. Lyon 2
http://preprints.cern.ch/archive/electronic/cern/preprints/thesis/thesis-99-060.pdf

[8] http://weblib.cern.ch/welcome.php

[9] Using Internet/Intranet Technologies in Library Automation / Martin Vesely, Jens Vigen (dir.). Genève : CERN , 2000. 67 p. Thèse : Univ. Economics Prague
http://documents.cern.ch/archive/electronic/cern/preprints/thesis/thesis-2000-040.pdf

[10] Contribution au développement d'un serveur de thèses électroniques / Carole Clerc, Jean-Michel Mermet (dir.). Lyon : INSA, 1999. 72 p. Rapport de stage : DESSID
http://www.enssib.fr/bibliotheque/documents/dessid/clerc.pdf

[11] http://documents.cern.ch/OAI

[12] The Santa Fe Convention of the Open Archives Initiative / Herbert Von de Sompel et Carl Lagoze. *In* : D-Lib Magazine, February 2000, vol. 6, n° 2
http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html