La gestion informatisée de corpus bibliographiques

Adaptation des normes et formats documentaires

L'interrogation de plusieurs banques de données (BDD) documentaires est une nécessité dans la constitution d'une bibliographie exhaustive sur un sujet pluridisciplinaire (14). Or, si la consultation de sources d'information variées, tant sur le plan géographique que thématique, garantit une meilleure couverture du sujet, elle pose néanmoins trois problèmes majeurs : l'hétérogénéité des formats de présentation ; la variabilité du vocabulaire d'indexation utilisé ; et la redondance de l'information due à la présence de références identiques.

Samuel Jolibois
INRS*
jolibois@ifrance.com

Emmanuel Nauer
LORIA**

Dominique Chouanière
Marc Mouzé-Amad
Françoise Grandjean
INRS

Jacques Ducloy
LORIA

* INRS : Institut national de recherche et de sécurité, Vandœuvre-lès-Nancy.
** LORIA : Laboratoire lorrain de recherche en informatique et ses applications, Vandœuvre-lès-Nancy.

L'utilisation d'un corpus documentaire par des chercheurs à des fins d'édition de références bibliographiques et d'analyse bibliométrique de leur domaine de recherche doit donc être précédée d'une phase de normalisation de la structure des notices ; d'une phase de normalisation du contenu des champs ; d'une phase de dédoublonnage des notices.

À chaque étape du traitement s'opèrent des choix qui peuvent être fondés sur des normes documentaires. Nous nous proposons ici de passer en revue les diverses normes existantes et d'en étudier leur adaptation à la gestion informatisée d'un corpus, alimenté et consulté par les chercheurs eux-mêmes. Seront exposées les solutions retenues dans le cadre de l'élaboration d'une base documentaire sur le *stress* professionnel à l'INRS (Institut national de recherche et de sécurité). Nous avons interrogé huit BDD sur cédérom. Trois proviennent du domaine biomédical : Medline et Biosis (États-Unis), EMBase (Pays-Bas). Trois autres sont spécialisées en sécurité et hygiène au travail : IOSHTIC (États-Unis), Cisilo (Suisse), INRS-B (France). Une BDD est spécialisée en psychologie, PsycLIT (États-Unis), et une autre est pluridisciplinaire, Pascal (France). Les équations de recherche ont été adaptées à chacune des bases interrogées, en raison de la diversité de leurs langages d'indexation respectifs. Au total, 27 000 notices bibliographiques ont ainsi été rassemblées.

Nous avons développé une application spécifique, dénommée WebStress, qui permet le reformatage des données et le dédoublonnage des notices, d'une part ; l'exploration hypertextuelle de la base documentaire dans une interface Web, d'autre part (30).

Normalisation de la structure des notices

Chaque base dispose d'une structure spécifique, constituée d'un nombre variable de champs. Ainsi, sur les huit BDD interrogées, nous avons recensé un total de 197 champs, ce qui fait une moyenne de 24 champs par base. Si certains sont communs à toutes les bases (*Auteurs, Titre, Date de publication*, etc.), d'autres sont spécifiques à une ou deux bases seulement (les *Tags*[1] par exemple, sont uniquement proposés par Medline, EMBase et PsycLIT). D'autres enfin apparaissent sous des formes différentes selon qu'ils sont fusionnés ou non. À titre d'exemple, le *Titre du périodique,* le *Volume,* le *Fascicule* et la *Pagination* apparaissent dans Medline PubMed dans un unique champ *Source*.

Pour uniformiser les données provenant de différentes bases, il est donc nécessaire de définir une structure respectant les normes documentaires en vigueur, ce qui suppose :
– la suppression des champs jugés non pertinents par rapport aux besoins des utilisateurs (Numéros de contrôle, Date d'entrée et de mise à jour dans la base originale, *CODEN*[2], *Cote, Langue du résumé, Public cible*, etc.) ;
– l'éclatement de certains champs (*Source* en : *Titre de périodique, Volume, Fascicule, Pagination, Lieu de publication, Éditeur, Date de publication*, etc. ; *Titre* en : *Titre original* et *Titre traduit*) ;
– la fusion de certains champs (*Descripteurs principaux* et *Descripteurs secondaires*, etc.) ;
– la création de nouveaux champs, absents de certaines bases (le champ *Type de publication* est absent de NIOSHTIC par exemple) ;
– la normalisation des intitulés des champs retenus (*Auteurs, Titre, Descripteurs*, etc.) ;

Normes de description bibliographique

Il existe deux types de normes de description bibliographique : les normes de catalogage et les normes éditoriales. Les premières sont très utilisées dans les grands centres documentaires (bibliothèques universitaires, par exemple) et nécessitent de fortes compétences en catalogage. Les secondes sont moins complètes, mais plus simples à utiliser ; elles sont donc adaptées à la documentation des chercheurs ou des petits centres d'information.

> Pour uniformiser les données provenant de différentes bases, il est nécessaire de définir une structure respectant les normes documentaires en vigueur

Normes de catalogage

L'ISBD-G (General International Standard Bibliographic Description), développée en 1975 par l'IFLA (International Federation of Library Associations and Institutions), est une norme internationale (18), dont dérivent toutes les règles de catalogage existantes de portée nationale : normes américaines (ANSI), françaises (AFNOR), etc. Cette norme de description bibliographique prescrit également la présentation des notices, y compris la ponctuation.

L'ISBD se décline en autant de normes spécifiques que de types de documents à décrire : monographies (ISBD-M), périodiques (ISBD-S), documents électroniques (ISBD-ER), etc. Elle comprend toutefois une structure générale en sept zones[3], qui vaut pour tous les documents, dont les trois premières seulement sont obligatoires :
– Zone 1. *Titre et mention de responsabilité* (titre original, titre traduit, auteur, affiliation de l'auteur) ;
– Zone 2. *Édition* (1re édition, 2e édition...) ;
– Zone 4. *Adresse bibliographique* (lieu de publication, éditeur, année de publication) pour les monographies ; *Numérotation* (volume, fascicule, numéro) pour les publications en série ;
– Zone 5. *Collation* (format du document, pagination, bibliographie, illustration...) ;
– Zone 6. *Collection* (titre et numéro dans la collection) ;
– Zone 7. *Notes* (notes sur les diverses parutions du document [reprint], les langues utilisées, etc.) ;
– Zone 8. *ISBN, reliure et prix* (monographies) ; *ISSN, titre clé et prix* (publications en série).

Les AACR2 (Anglo-American Cataloguing Rules 2nd ed.), parues en 1978 et révisées en 1988 (3), incorporent les règles de l'ISBD, tout en conservant leurs spécificités et en intégrant les besoins nouveaux, nés de l'informatisation des bibliothèques. Elles ont une portée interna-

1. Les tags sont des descripteurs génériques qui renseignent sur le type de population (mâle, femelle, animal...) ou la tranche d'âge (enfant, adulte...) concernée, le type de publication (étude de cas, étude de suivi...), l'aire géographique...
2. Le CODEN est un code alphanumérique d'identification des périodiques, qui tend à disparaître au profit de l'ISSN.

3. La zone 3 n'est pas attribuée pour les monographies, mais peut être utilisée pour les autres types de documents (cartes et plans par exemple).

tionale et ont été traduites et adaptées en plusieurs langues, notamment le français et l'espagnol.

Normalisation éditoriale

La norme internationale ISO 690 (21) définit la présentation des références bibliographiques[4]. Elle ne reprend que les éléments les plus importants de la notice bibliographique de l'ISBD puisqu'elle n'est pas destinée aux bibliothécaires, mais aux chercheurs et rédacteurs de bibliographies.

L'ISO 690 décrit des champs obligatoires (*Auteurs, Titre, Année de publication*, etc.) et facultatifs (*Pagination, Collection, Notes*, etc.), mais laisse une certaine liberté au niveau de la présentation des éléments bibliographiques retenus.

Il faut alors appliquer les normes éditoriales reconnues à l'intérieur de sa discipline[5] : Chicago Style (histoire), MLA Style (Modern Language Association, arts et lettres), APA Style (American Psychological Association, psychologie), CBE Style (Council of Biological Editors, biologie), AMA Style (American Medical Association, médecine), Vancouver Style (médecine). Ces trois derniers formats de présentation sont très proches, mais c'est le dernier (plus connu en France sous le nom de Convention de Vancouver) que nous avons retenu[6], parce qu'il fait autorité dans le domaine médical. En effet, ce format est adopté par les 500 plus grandes revues biomédicales. Le style de Vancouver (16) reprend dans ses grandes lignes les recommandations de l'*Index Medicus* (42), qui est l'édition papier de la base bibliographique Medline, produite par la NLM (National Library of Medicine). Les autorités (auteurs, descripteurs, titres de périodiques) doivent notamment être identiques à celles de l'*Index Medicus*.

Formats d'échange de données

Une fois définie la structure de la notice bibliographique se pose le problème de sa représentation informatique. Il existe plusieurs formats normalisés de codage de données pour l'échange électronique d'information. Nous nous limiterons à la présentation des deux formats les plus utilisés en documentation : MARC et SGML. Nous aborderons brièvement la TEI (Text Encoding Initiative) et le Dublin Core Metadata Element Set, deux normes qui ont tendance à se développer avec la croissance des documents électroniques.

> MARC est un format normalisé, destiné à la gestion et à l'échange de données bibliographiques informatisées

MARC

MARC (Machine Readable Cataloguing) est un format normalisé destiné à la gestion et à l'échange de données bibliographiques informatisées. Le format MARC original a été développé par la Library of Congress en 1966 (33), tandis que la British Library travaillait parallèlement à l'établissement de son propre format MARC. Ce sont donc deux versions qui ont vu le jour : USMARC et UKMARC.

Après trente années d'existence, MARC est le format de codage des données bibliographiques le plus utilisé en documentation. Il fait l'objet d'une norme internationale établie en 1973, révisée en 1996, [ISO 2709] (24), mais a subi des adaptations nationales, à tel point que l'on compte aujourd'hui une cinquantaine de formats MARC spécifiques[7], utilisés dans plus de soixante pays (9).

Un format d'uniformisation, l'UNIMARC[8], a vu le jour en 1972, fondé sur des normes internationalement reconnues : l'ISO 2709 (24) déjà citée, l'ISO 3166 (25) (codes de pays), l'ISO 639 (20) (codes de langues), l'ISBN-ISO 2108 (23), l'ISSN-ISO 3297 (26) et l'ISBD (règles de catalogage). Actuellement, seules douze agences bibliographiques nationales[9] ont adopté UNIMARC, le format américain USMARC fait donc toujours référence.

MARC est basé sur une structure de données comportant :
– des champs de longueur fixes, utilisés pour stocker des données de

4. La norme ISO 690-2 :1997 s'intéresse plus spécifiquement à la description bibliographique des documents électroniques.
5. Une présentation de ces formats est disponible en ligne sur le site de Bedford Books (http://www.bedfordbooks.com/rd/) et sur celui de l'université de Washington (http://healthlinks.washington.edu/hsl/styleguides/).
6. La bibliographie placée à la fin de cet article est formatée selon le style de Vancouver.
7. Une trentaine de ces formats sont basés sur USMARC, d'autres sur UKMARC, d'autres encore sont hybrides.
8. Une description détaillée d'UNIMARC est disponible en ligne : IFLA (International Federation of Library Associations and Institutions), UNIMARC: an Introduction. [online]. 1996. [cited 1998 Nov 20]. Available from URL http://ifla.inist.fr/VI/3/p1996-1/unimarc.htm
9. Le décret n° 34-174 du 5 février 1993 (JO du 07/02/1993) a confirmé le rôle d'UNIMARC comme format national d'échange bibliographique en France. UNIMARC est notamment utilisé par la Bibliothèque nationale de France.

nature générale, numéros d'identification *(ISBN, ISSN, Numéro de la bibliographie nationale*) ou valeurs codées (*Pays*, *Langue*) ;
– des champs de longueur variable, où sont enregistrées les données bibliographiques spécifiques (*Titre, Mention de responsabilité*, etc.) ;
– des sous-champs (le champ *Adresse bibliographique* par exemple se décompose en : *Lieu de publication, Éditeur, Date de publication*).

Tous les champs sont délimités par des identificateurs numériques, qui ne facilitent pas la lisibilité du format (par exemple, 010=*ISBN*, 101=*langue du document*, 410=*collection*, 700=*nom de personne*, etc.). Les sous-champs sont identifiés par des lettres (par exemple, 400$a=*adresse bibliographique*, 400$b=*lieu de publication*, 400$c=*date de publication*).

Le format MARC est complexe à utiliser et manque de convivialité (15). S'il est adapté aux grands centres documentaires disposant d'un personnel formé, nous ne l'avons pas jugé pertinent pour représenter nos données.

SGML

La norme SGML (28) se caractérise par sa capacité à représenter, à l'aide de balises ou identificateurs, la structure logique de n'importe quel document, indépendamment de la plate-forme de travail. Chaque élément (titre, sous-titre, image, légende, note de bas de page, paragraphe, etc.) est identifié par une balise de début et de fin d'élément.

```
<notice>
        <auteur>
                <e>premier auteur</e>
                <e>second auteur</e>
                <e>...</e>
        </auteur>
        <titre>titre</titre>
        <date>date</date>
</notice>
```

La norme SGML se développe rapidement dans le monde de la documentation et va peu à peu se substituer au format MARC (15, 35). Nous l'avons retenue ici pour les avantages suivants :
– intégration de formats différents (MARC[10] ou non-MARC) par adaptation des balises SGML à ces formats ;
– lisibilité et compréhensibilité ;
– recherche structurée sur tous les éléments du document ;
– adaptabilité rapide au format HTML pour la mise en place d'une navigation hypertextuelle ;
– adaptabilité aux données bibliographiques qui comprennent du texte et sont structurées ;
– indépendance de la plate-forme et du logiciel d'édition ou de consultation, d'où une réutilisation facilitée des données.

TEI P3

La norme TEI P3 (Text Encoding Initiative Public Proposal Number 3), parue en 1994 (13), constitue une application particulière de SGML. L'objectif de la TEI est de définir un format standard de codage et d'échange de textes électroniques. La norme TEI P3 définit un ensemble de conventions de balisage et d'annotation des textes conforme à la norme SGML.

La TEI couvre toute forme de publication (dictionnaires, manuscrits, corpus linguistiques, données terminologiques, etc.). Le chapitre 6 de la norme est consacré aux citations et références bibliographiques et précise quelles sont les balises à retenir pour la description bibliographique d'un document, quels intitulés leur donner et comment les présenter. Le schéma proposé obéit au cadre général de description bibliographique posé par l'ISBD et offre une totale compatibilité avec les champs retenus par les logiciels de gestion de références bibliographiques très utilisés dans le monde de la recherche comme BibTeX, EndNote, Reference Manager et Procite. Il existe également une table de correspondance entre les champs du format MARC et les balises de la TEI P3.

10. La Library of Congress propose une DTD (Document Type Definition) pour effectuer l'intégration du format MARC en SGML [LC 1998].

Dublin Core Element Set

Le Dublin Core Metada Element Set, établi en 1995 par l'OCLC (37) et le NCSA (National Centre for Supercomputer Applications), se démarque des formats précédents en ce qu'il propose un ensemble minimal de description bibliographique des documents électroniques uniquement. Cet ensemble se compose de quinze champs seulement : titre, créateur, sujet, description, éditeur, contributeurs, date, type de ressource, format, identificateur de ressource, relation, localisation, droits. Ainsi, en raison de l'utilisation aisée du format Dublin Core, ces champs peuvent être renseignés par les auteurs de documents électroniques eux-mêmes.

Normalisation du contenu des champs

Une fois les structures de notices bibliographiques uniformisées, il est nécessaire de normaliser le contenu des champs les plus importants pour la recherche documentaire et l'exploitation bibliométrique du corpus : *Date de publication, Langue de publication, Type de publication, Titre de périodique, Pays d'affiliation de l'auteur principal, Auteurs, Affiliation de l'auteur principal* et *Descripteurs*. Pour la normalisation de ces champs dans notre corpus, nous avons suivi les recommandations de Vancouver ou,

à défaut, celles de l'*Index Medicus*. Lorsqu'il existe une norme internationale ISO, nous signalons sa présence.

Normalisation des dates

Nous n'avons retenu dans le champ *Date de publication* que l'année de publication. Lorsque plusieurs années étaient présentes, nous avons retenu la plus récente. La mention du mois ou du jour a été supprimée. En revanche, la date d'un congrès apparaît en entier sur le modèle anglo-saxon : année, mois, jour (1998 Dec. 12). Ce schéma est conforme à la norme internationale ISO 8601 (27) sur l'écriture des dates et des chronologies, qui annule et remplace les normes ISO 2014, ISO 2015, ISO 2711, ISO 3307 et ISO 4031.

Normalisation des langues

La mention de la langue de publication apparaît sous différentes formes selon la base interrogée. Nous avons retenu l'intitulé complet de la langue proposé dans la norme internationale (20). L'intitulé nous semble plus lisible que le code terminologique en deux lettres ou le code bibliographique en trois lettres.

Normalisation des types de documents

Les types de document ont des dénominations distinctes d'une base à l'autre. Dans notre corpus par exemple, les actes de congrès apparaissent sous quinze dénominations différentes : « Meeting Report », « Conference Presentation », « Conference-paper », « MEETING PAPER », « MEETING POSTER », « MEETING REPORT », « MEETING SLIDE », « Meeting Summary », « Meeting-Document », etc. Dans notre exemple, nous avons éliminé ces variations en retenant les dénominations de types de publication figurant dans l'*Index Medicus*.

Normalisation des titres de périodiques

Les titres de périodiques apparaissent selon les bases interrogées, tantôt dans leur forme complète, tantôt dans leur forme abrégée. Des variations peuvent en outre exister dans la transcription des titres étrangers (traduction ou translittération).

Pour normaliser les titres de périodiques de notre corpus, nous avons retenu comme forme normale l'édition 1998 des publications en série (8 866 titres) dépouillées par la NLM pour ses BDD bibliographiques[11], comme le préconise la Convention de Vancouver. Cette liste recense les publications analysées dans l'*Index Medicus*, l'*Index to Dental Literature*, l'*International Nursing Index* et l'*Hospital and Health Administration Index*. À défaut, nous avons retenu dans cet ordre de priorité les 8 379 périodiques indexés par l'ISI[12] (*Current Contents*), les 1 409 périodiques analysés par PsycLIT[13], les 1 065 périodiques exclusivement dépouillés dans EMBase[14] et les 151 titres couverts par NIOSHTIC[15]. Enfin, pour les périodiques n'apparaissant dans aucune des listes précitées, nous avons respecté les règles fixées par la norme internationale ISO 4 (19) qui donne les règles d'abréviation pour les mots du titre ou du titre de publication.

Nous avons conservé la version abrégée et complète des titres de périodiques, afin de favoriser les recherches documentaires et l'édition de listes bibliographiques personnalisées.

Normalisation des pays

Les noms de pays subissent également de nombreuses variations. Ainsi, les États-Unis peuvent être désignés par USA, États-Unis, United States, US. Les changements géopolitiques intervenus ces dernières années compliquent les dénominations : éclatement de l'URSS en plusieurs républiques indépendantes, éclatement de la Yougoslavie, de la Tchécoslovaquie, changement de noms de pays africains, etc. Il existe une norme internationale ISO 3166 (25) fixant la dénomination des pays et proposant un code à deux ou trois chiffres (par exemple, France, FR, FRA).

Conformément à la Convention de Vancouver, nous avons retenu pour notre corpus les dénominations des pays telles qu'elles apparaissent dans le MeSH *(Medical Subject Headings)*, le thésaurus de Medline, qui présente l'avantage d'un classement géographiquement hiérarchisé : continent, région, pays, état, ville. Nous avons laissé inchangées les entités géographiques lorsqu'il était impossible d'identifier une subdivision. Par exemple, nous avons laissé l'entité Yugoslavia, sauf si un nom de ville figurait nous permettant d'identifier le pays (par exemple : Zagreb Croatia, Beograd Serbia).

Normalisation des auteurs

La description des auteurs subit de nombreuses variations (8). Le nom précède généralement le prénom,

11. National Library of Medicine, List of Serials Indexed for Online Users. [online]. [Cited 1998 Nov 12]. Available from URL ftp://nlmpubs.nlm.nih.gov/online/journals/
12. ISI Master Journal Coverage List. [online]. [Cited 1998 Nov 12]. Available from URL http://www.isinet.com/listlinx.html
13. PsycInfo and PsycLIT Journal Coverage List. [online]. [Cited 1998 Nov 12]. Available from URL http://www.apa.org/psycinfo/covlist.html
14. List of Journals Only to Be Found in EMBase. [online]. [Cited 1998 Nov 12]. Available from URL http://www.chest.ac.uk/datasets/embase/journals.html
15. NIOSHTIC Core Journal List. [online]. [Cited 1998 Nov 12]. Available from URL http://www.cdc.gov/niosh/nioshtic.html

Quelques variations dénominatives de l'École de management de l'université de Manchester
(qui comptabilise 34 dénominations différentes dans notre corpus)

- Victoria U of Manchester, Inst of Science & Technology, England
- Victoria U of Manchester Inst of Science & Technology, Manchester School of Management, England
- Manchester Sch. Management, Univ.
- Manchester Inst. Sci. and Technol., Manchester, U.K.
- Manchester School of Management, UMIST
- Manchester School of Management, University of Manchester, Institute of Science and Technology
- School of Management, Institute of Science and Technology, University of Manchester, PO Box 88, Manchester M60 1QD, United Kingdom
- U Manchester Inst of Science & Technology, Manchester School of Management, Manchester, England

mais parfois c'est l'inverse. Le prénom est complet ou abrégé, ou bien seul le premier prénom apparaît en entier et le second est abrégé. Les noms peuvent être composés et présenter des particules (de, von, van, etc.). À ces variations s'en ajoutent d'autres, comme celles liées à des séparateurs (tiret, apostrophe, virgule) ou à la casse. Par exemple, l'auteur le plus référencé dans notre corpus documentaire sur le *stress* professionnel, Cary L. Cooper apparaît sous neuf formes différentes « C. L. Cooper », « COOPER CL », « COOPER-C-L », « Cooper C.L. », « Cooper CL », « Cooper, C. L. », « Cooper,-Cary-L. », « Cooper-C-L », « Cooper-CL ».

Les règles de catalogage (3) et (1) préconisent le rejet des particules selon les pratiques du pays auquel appartient l'auteur. Ainsi, un auteur français, espagnol ou portugais verra sa particule « de » rejetée (par exemple : Roux, Jean de) tandis qu'un auteur américain ou un italien la conservera en tête (De Sicca, Giovanni). Les noms composés sans espace sont classés au premier élément lorsqu'il s'agit de Français, d'Allemands, d'Espagnols (par exemple, Garcia Lorca, Federico), au dernier élément lorsqu'il s'agit d'Anglo-saxons (par exemple, Mill, John Stuart) ou de Portugais (par exemple Antunes, António Lobo).

Ces règles sont difficilement automatisables. Aussi avons-nous adopté un format plus simple qui ne procède à aucun rejet. Le nom complet de l'auteur apparaît suivi des initiales de ses prénoms, sans point, conformément à l'usage de l'*Index Medicus*. Les noms composés sont classés au premier élément et les éléments sont systématiquement séparés par un trait d'union, lorsqu'ils n'en possèdent pas.

Normalisation de l'affiliation des auteurs

Le champ *Affiliation de l'auteur* désigne l'organisme dans lequel travaille le premier auteur de la publication. Il est composé de différentes parties (section, division, département, unité, adresse, ville, code postal, pays, etc.) dont le nombre et l'ordre subissent de fortes variations.

Pour chacune des parties, les variations de présentation sont liées à l'utilisation de sigles (par exemple, UMIST University of Manchester Institute of Science and Technology) et d'abréviations (par exemple, University se retrouve sous cette forme ou sous des formes abrégées « Univ. », « Univ », « U »), ou encore à la présence ou l'absence de mentions relatives à un statut commercial (limited, ltd, incorporated, inc, Gmbh, SA, etc.). On retrouve également les variations liées aux dénominations des noms de pays (*cf.* normalisation des pays).

L'ensemble de ces variations produit une explosion combinatoire dans la présentation des affiliations, comme nous pouvons le constater dans l'encadré ci-dessus.

Les règles de catalogage (3, 2) préconisent de conserver l'écriture la plus courante de l'organisme. Ainsi, le nom de la collectivité doit être saisi dans la langue utilisée couramment par cette collectivité dans ses publications, sous leur forme développée de préférence au sigle, sauf si celui-ci est la forme la plus connue. En général, c'est la collectivité subordonnée qui doit apparaître en premier (par exemple, Institute of Occupational Medicine and Health, University of Rochester), mais, dans la pratique, ce principe est sujet à de nombreuses variations.

Au total, les normes documentaires officielles sont conçues comme un ensemble de règles, contredites par de nombreuses exceptions. Ces règles laissent une trop large part à l'interprétation pour être automatisées. Nous avons donc conçu nos propres règles, plus conformes à un traitement automatique. La première consiste à développer tous les sigles et abréviations (Univ University, Dpt Department, Sch School, etc.) sauf s'ils sont d'usage universel (UNESCO, ONU, OTAN, etc.). La seconde consiste à identifier les différentes parties de l'affiliation pour les ordonner, du général vers le particulier. Ainsi, nous aurons par exemple : United

Quelques statistiques sur le corpus bibliographique traité

26 251 notices initiales retrouvées, sur la période* 1967-1997
20 402 notices après dédoublonnage, soit 5 849 doublons éliminés

1.1 Ventilation du nombre de notices par base après dédoublonnage

Base	Notices
Medline	8 004
EMBase	3 125
Biosis	1 446
PsycLIT	4 137
Pascal	1 252
NIOSHTIC	1 807
Cisilo	498
INRS-B	132
Total	**20 401**

1.2 Ventilation du nombre de doublons par base

Base	Doublons
Medline	53
EMBase	991
Biosis	1 167
PsycLIT	1 411
Pascal	822
NIOSHTIC	1 062
Cisilo	233
INRS-B	110
Total	**5849**

2. Problèmes rencontrés

Les règles préconisées par les normes et formats documentaires en vigueur ne sont pas toujours complètement automatisables. Les principaux problèmes qui demeurent concernent les champs Auteurs, Affiliation des Auteurs, Source, et Descripteurs.
Notre algorithme de reformatage du champ Auteurs fonctionne bien dans l'ensemble. Il adopte, rappelons-le, le format Nom Initiales des prénoms. Certaines ambiguïtés n'ont toutefois pas pu être levées actuellement. C'est le cas notamment pour les auteurs qui possèdent le même nom, mais dont l'une des initiales varie.
Par exemple, l'auteur Robert A. Karasek apparaît, après normalisation, sous les formes « Karasek RA » et « Karasek R ». Nous savons qu'il s'agit du même auteur, mais comment le faire reconnaître par le système. Une des solutions envisagées (non encore testée) consiste à vérifier qu'ils appartiennent au même organisme. Mais nous nous heurtons alors au problème de l'Affiliation.
Nous n'avons actuellement pas encore trouvé d'algorithme satisfaisant pour traiter les variations du champ Affiliation des auteurs. Considérons les deux variations suivantes concernant le service Physiologie environnementale de l'INRS. Comment faire reconnaître par le système l'identité entre « INRS Nancy » et « INRS » ainsi que celle entre « Cent. res. » et « cent. rech. » ?

Cent. res. INRS Nancy, serv. physiologie environmentale, Nancy, FRA
INRS, cent. rech., serv. physiologie environnementale, Nancy, FRA

Notre algorithme de reformatage du champ Source fonctionne actuellement uniquement pour les articles de périodiques, qui représentent 90 % de notre corpus. Nous n'avons pas encore trouvé de solution satisfaisante pour normaliser la source des non-périodiques, tant les cas de figure sont divers. Outre la diversité des types de documents (monographie, contribution à une monographie, chapitre d'une monographie, colloque, acte d'un colloque, rapport, etc.), nous nous heurtons à la variation dans le catalogage de ces documents par les différentes bases de données interrogées.
La normalisation automatique des descripteurs passe par la mise en correspondance des termes de notre corpus avec ceux d'un métathésaurus. À cet effet, nous avons utilisé l'UMLS (Unified Medical Language System), riche de quelque 476 000 concepts, représentant plus d'un million de formes différentes prélevées parmi 71 sources différentes. Malgré cette richesse, inégalée dans le domaine biomédical, plus de 8 000 termes de notre corpus n'ont pas été reconnus, parmi les 17 000 termes obtenus après reformatage (sur un nombre initial de 45 000). Il reste donc à traiter ce reliquat manuellement.

3. Projets de développement de l'application WebStress

La méthode et l'application décrite ici peuvent être réutilisées pour traiter tout corpus bibliographique dans le domaine biomédical, puisque nous avons retenu une norme éditoriale adoptée par des éditeurs de revues médicales.
Il est possible de traiter des corpus bibliographiques dans d'autres domaines (histoire, droit, informatique, etc.), moyennant la modification des filtres de reformatage afin qu'ils correspondent à la norme éditoriale du domaine considéré.
WebStress n'est pas seulement une application de reformatage de données et de dédoublonnage de notices. C'est également une interface Web de consultation hypertextuelle du corpus bibliographique. L'interrogation peut se faire à partir d'une liste d'items (auteurs, descripteurs) classés par fréquence décroissante ou à partir de requêtes booléennes classiques.
Une fonction de classification automatique par clusterisation permet de dégager des réseaux d'auteurs (les « collèges invisibles ») ainsi que des thèmes de recherche spécifiques au sein d'une problématique plus vaste (dans notre exemple, le stress professionnel).
L'originalité de cette interface est qu'elle permet de croiser les informations issues de données locales (le corpus) et de données distantes (BDD ou thésaurus disponibles sur Internet). Nous pouvons par exemple exporter une requête appliquée en premier sur le corpus local vers une BDD comme Medline PubMed ou un moteur de recherche sur Internet.
Nous envisageons dans le futur d'interfacer l'application WebStress avec des logiciels de bibliographie tels que Reference Manager, EndNote ou Procite, afin de profiter de leurs fonctionnalités très riches en matière de publication de bibliographies à destination de périodiques spécialisés. Ainsi disposerons-nous d'une chaîne complète de traitement de l'information bibliographique, depuis l'interrogation des BDD elles-mêmes jusqu'à la publication de bibliographies.

* La période couverte varie selon la disponibilité des BDD interrogées sur cédérom : Medline (1967-1997), EMBase (1984-1997), Biosis (1990-1997), PsycLIT (1983-1997), Pascal (1987-1997), NIOSHTIC (1973-1997), Cisilo (1974-1997), INRS-B (1981-1997). Nous procéderons à la mise à jour du corpus avant la fin de l'année.

Kingdom, University of Manchester, Institute of Science and Technology, School of Management.

Cette normalisation permet un meilleur repérage des affiliations des auteurs, en proposant un classement à plusieurs niveaux hiérarchiques. Ainsi pourrons-nous étudier les pays, villes et universités les plus productives sur un sujet déterminé ; les instituts spécialisés dans un domaine, toutes universités confondues ; etc.

Normalisation des descripteurs

Les principales variations terminologiques observées dans notre corpus sont d'ordre morphologique (« Work Load », « Work-Load », « Workload ») et lexical (« Corticotropin », « ACTH », « Acth Hormone », « Adrenocorticotrophic Hormone », « Adrenocorticotropic Hormone »).

Conformément aux recommandations de Vancouver, nous avons utilisé le MeSH et le metathesaurus UMLS (Unified Medical Language System, produit par la NLM) (40, 31) pour la normalisation des descripteurs. Sur le plan informatique, la normalisation s'est opérée de la manière suivante (39) :

– réduction des mots-clés du corpus à une forme minimale, expurgée des caractères non alphanumériques ; les majuscules ont été converties en minuscules ;

– comparaison avec les termes équivalents contenus dans l'UMLS et sélection du terme préférentiel en cas de concordance ;

– lemmatisation simple des mots-clés non reconnus dans l'UMLS, afin d'éliminer les variations morphologiques (marques du pluriel, américanismes) ;

– en cas de non-concordance avec les termes de l'UMLS, la forme la plus occurrente devient la forme préférentielle.

Dans notre application, ce type de traitement informatique a permis d'éliminer, parmi un corpus de 45 000 descripteurs, plus de 60 % de formes concurrentes. Les descripteurs ont ensuite subi une analyse sémantique par un expert en vue de la construction d'un thésaurus sur le *stress* professionnel : établissement de relations d'équivalences entre les descripteurs (termes préférentiels) et les non-descripteurs (formes concurrentes), établissement des relations hiérarchiques (génériques, spécifiques) et associatives.

Dédoublonnage des notices

Les doublons désignent toutes les notices, au sein d'une ou de plusieurs BDD, qui font référence à la même publication logique : auteurs, titres et support de publication[16] identiques. L'élimination des doublons est une opération préalable à toute analyse bibliométrique d'un corpus. Dans notre application, la présence de 6 000 références en double au sein de la base *Stress* faussait les résultats statistiques et ne permettait pas l'édition de listes bibliographiques rigoureuses.

Quelle que soit la solution informatique retenue, le dédoublonnage s'opère en trois étapes (11) : construction d'une ou de plusieurs clés de dédoublonnage ; identification des doublons par comparaison des clés de dédoublonnage (mise en correspondance) ; élimination des doublons du corpus.

Construction de la clé de dédoublonnage

La sélection des champs qui vont servir à la construction de la clé de dédoublonnage s'avère cruciale. Il convient de sélectionner les champs qui sont présents dans toutes les notices. Il est également essentiel que ces champs possèdent un format homogène, soient significatifs et identifient de façon univoque les notices. C'est pourquoi la normalisation est une étape critique dans la construction de cette clé. Deux solutions sont envisageables : le recours à des codes normalisés d'identification d'unités d'information et la création d'une clé de dédoublonnage spécifique.

Systèmes d'identification univoque d'unités d'information

Il existe plusieurs systèmes internationaux d'identification univoque d'unités d'information (42, 36). Certains s'attachent à identifier un ensemble de documents (périodique, ouvrage collectif, conférences, etc.) ; on retrouve ici le CODEN et l'ISSN pour les périodiques, l'ISBN pour les ouvrages. D'autres, comme le SICI ou le BIBLID (*cf.* ci-après), portent sur l'identification du document lui-même (article de périodique, contribution à une monographie, contribution à une conférence...). D'autres enfin décrivent une partie de document (schéma, tableau, bibliographie, etc.) ; on a alors recours aux notices MARC, aux URL (Uniform Resource Locator) ou encore aux DOI (Digital Object Identifier).

Dans le but d'identifier les doublons – lors d'une interrogation multibase – nous nous sommes intéressés plus particulièrement aux systèmes qui permettent de caractériser les documents eux-mêmes, à savoir : BIBLID et SICI.

16. On observe parfois des variations (erreurs orthographiques, omissions, etc.) concernant les auteurs ou le titre d'une même publication, d'une base à l'autre. Nous ne considérons pas comme des doublons les documents écrits par le même auteur et portant le même titre, mais qui ont fait l'objet de publications différentes (par exemple, un rapport interne qui devient un article de périodique).

– BIBLID (BIBLiographic IDentification) est une norme internationale ISO 9115 (29) établie en 1987, mais qui n'a plus cours depuis 1996 en raison de son inadéquation aux caractéristiques des publications électroniques. Cet identifiant retient, pour les articles de périodiques, l'ISSN, l'année de publication, la tomaison (volume, numéro, partie) et la pagination ; pour les contributions à des ouvrages, l'ISBN, l'année de publication, la pagination. Les zones sont précédées de la mention BIBLID et séparées par des signes de ponctuation.

Exemple :

BIBLID 0272 17716(1983)3 : 3p.68-70
BIBLID3-8007-1317-9(1986)p.158-170

– Le SICI (Serial Item and Contribution Identifier), défini en 1991 par le SISAC (US Serials Industry Standardization Advisory Committee), a fait l'objet d'une normalisation américaine (4). Le SICI identifie de façon unique chaque expression physique d'une même entité logique (plusieurs publications d'un même article auront des SICI différents). Une version révisée du SICI est parue en 1996, qui offre de nouvelles spécificités, comme la désignation du support (électronique par exemple) ou les composantes d'un document. Le SICI s'avère peu aisé à construire et à manipuler (il comporte quarante caractères, calculés à partir de l'ISSN, de la tomaison, de l'année de la publication, etc.).

Nous n'avons pas retenu ces deux identificateurs comme clés de dédoublonnage, bien qu'il s'agisse de normes, parce qu'ils ne nous semblent pas adaptés à notre corpus. En effet, la mention de l'ISSN, notamment, ne figure pas dans les bases INRS-B et NIOSHTIC. De plus, la tomaison, malgré le reformatage, est sensible à certaines variations (supplément, numéro spécial, etc.). Nous avons donc construit une clé de dédoublonnage adéquate.

Construction d'une clé de dédoublonnage spécifique

La clé de dédoublonnage nécessite l'extraction d'une nouvelle information, à partir du traitement et de la concaténation de plusieurs données, sous une forme normalisée. Nous nous sommes inspirés du code Meyer-Uhlenried (32). Cette clé alphanumérique de treize caractères comprend les quatre premières lettres du nom de l'auteur, les initiales des deux premiers prénoms de l'auteur si disponibles, les deux dernières lettres de l'année, la première lettre des cinq premiers mots du titre.

Nous avons apporté quelques aménagements à ce code :
– l'indication de la première page de l'article a été ajoutée, ce qui évite tout risque de confusion entre deux articles de périodique (c'est par exemple le cas des articles en plusieurs parties) ;
– les quatre lettres de l'année ont été retenues pour éviter toute ambiguïté (passage à l'an 2000) ;
– lorsque le titre comporte moins de cinq mots, nous avons retenu les initiales des mots présents que nous avons complétées par ajout des lettres du dernier mot du titre afin d'obtenir un code de cinq caractères ;
– nous avons constitué deux clés de dédoublonnage, la première (Dub1) fonctionnant d'après le titre original, la seconde (Dub2) d'après le titre traduit, ceci pour multiplier les chances de retrouver des doublons (certaines bases telles Biosis ou NIOSHTIC ne proposent pas de titre dans sa langue originale) ;
– pour le codage du titre, seules les lettres ont été conservées. Pour l'ensemble de la clé, tous les caractères non imprimables de même que les tirets, points, apostrophes, etc., ont été supprimés ; chaque élément du code est séparé du suivant par une étoile et toutes les lettres ont été converties en majuscules ;
– dans le cas des non-périodiques, nous avons retenu le nombre total de pages du document (à la place de la première page), qui est facilement identifiable grâce à la mention « p. » (par exemple, 230 p.) ; lorsqu'il s'agit d'une contribution à une monographie, c'est l'indication de la première page de la contribution qui est sélectionnée et non la pagination totale.

Exemple :

Tamburro GA.; Di-Sciascio G; De-Giglio F. Les facteurs déontologiques du personnel psychiatrique en tant que symptôme d'un état de «Burn-out». [Infringements of the ethics of the psychiatric profession: A Burn-out syndrome]. Psychologie-Médicale1992;
24 (Spec Issue 4):372-376.

Dub1: TAMB*GA*92*LFDDP*372
Dub2: TAMB*GA*92*IOTEO*372

Élimination des doublons

La double clé de dédoublonnage (titre original et titre traduit) permet d'identifier de manière univoque une publication. Ainsi deux ou plusieurs notices présentant la même clé de dédoublonnage sont considérées comme des doublons.

En présence de doublons, il convient de ne conserver qu'une seule notice et d'éliminer les autres. La sélection des notices à conserver obéit à un ordre de priorité dépendant de la base d'origine. Nous retenons dans cet ordre la notice provenant de Medline, puis celle d'Embase, de Biosis, de PsycLIT, de Pascal, de NIOSHTIC, de Cisilo, et enfin celle d'INRS-B.

Conclusion

Les problèmes de format sont familiers aux professionnels de la documentation et des bibliothèques qui ont conçu et utilisent des systèmes de description bibliographique élaborés. Il est difficile de

faire un choix parmi les nombreuses normes documentaires existantes, qu'il s'agisse des normes de catalogage comme des formats d'échange de données (10, 38).

Nous dressons trois constats de notre revue des normes documentaires et de son adaptation à notre application :

– le caractère multiforme des normes, même les mieux établies ; le meilleur exemple en est la norme MARC qui revêt autant de formes que de pays l'utilisant, soit une cinquantaine recensée par l'IFLA (17) ;

– l'inadéquation des normes existantes pour les traitements informatisés ; les règles de catalogage ont été conçues initialement pour être utilisées par un opérateur humain qui dispose du document original et non par un ordinateur, d'où une grande liberté dans l'interprétation de ces règles, inacceptable dans le cadre d'une solution informatique ;

– l'inadéquation des identificateurs univoques d'unités d'information (BIBLID ou SICI) pour le dédoublonnage des notices ; les notices bibliographiques issues d'une interrogation multibase ne possèdent pas tous les éléments requis par ces identificateurs, d'où la nécessité de construire une clé de dédoublonnage adéquate.

En réponse à la variété de normes de catalogage, tend à s'imposer la notion de *core record* (12), qui désigne l'élaboration d'une notice minimale, ne retenant que les champs essentiels à l'identification bibliographique d'un document. C'est le parti que nous avons suivi, en nous fondant sur les recommandations de la Convention de Vancouver.

L'absence du document original est un réel obstacle à la description bibliographique des documents, notamment dans l'écriture du nom de l'auteur ou de son affiliation. Or, toutes les normes, y compris la Convention de Vancouver, stipulent le recours au document original dans la rédaction d'une liste bibliographique : « *The references must be verified by the author against the original documents* ». Cet obstacle risque de devenir plus prégnant dans les années à venir, avec la généralisation des documents électroniques et l'accroissement de la littérature scientifique.

L'inadéquation des normes documentaires (ISBD, AACR2) à un traitement informatique est relevée par les spécialistes, qui dénoncent la « *dichotomie entre les règles de catalogage et les formats informatiques utilisés pour présenter les informations* » (12). C'est pourquoi nous avons souvent opté pour une solution originale, adaptée à l'exploitation bibliométrique du corpus, parfois éloignée des normes documentaires en vigueur, comme pour les noms d'auteurs ou de collectivités-auteurs, en attendant la parution de nouvelles normes, plus adaptées au traitement automatique des données bibliographiques.

Septembre 1999

BIBLIOGRAPHIE

Cette bibliographie est formatée selon le style de Vancouver.

1. AFNOR NF Z44-001 :1995. Technologies de l'information - Classement alphabétique des dénominations. Paris : AFNOR; 1995.

2. AFNOR NF Z44-060 :1983. Documentation - Catalogue d'auteurs et d'anonymes - Forme et structure des vedettes des collectivités-auteurs. Paris : AFNOR; 1983.

3. Anglo-American Cataloguing Rules (AACR2). 2nd ed., rev. Chicago : American Library Association; 1988.

4. ANSI/NISO Z39-56 : 1996 (revision of 1991). Serial Issue and Contribution Identifier (SICI). Bethesda : NISO Press; 1996.

5. Bedford Books. The Humanities: MLA Style. [online]. 1997. [cited 1998 Nov 20]. Available from : URL : http://www.bedfordbooks.com/rd/ctmla.html

6. Bedford Books. The Sciences: CBE Style. [online]. 1997. [cited 1998 Nov 20]. Available from : URL : http://www.bedfordbooks.com/rd/ctcbe.html

7. Bedford Books. The Social Sciences: APA Style. [online]. 1997. [cited 1998 Nov 20]. Available from : URL : http://www.bedfordbooks.com/rd/ctapa.html

8. Degez, Danièle. Compatibilité des langages d'indexation. Mariage, cohabitation ou fusion ? Quelques exemples concrets. *Documentaliste-Sciences de l'information*, 1998;35(1):3-14.

9. Delsey, T. L'Évolution des formats MARC. In : L'Avenir des formats de communication. Banque internationale d'information sur les États francophones de l'ACCT; 1996 Oct 7-11; Bibliothèque nationale du Canada, Ottawa.

10. Dempsey, Lorcan ; Mumford, Anne ; Robiette, Alan, et al. eLib standards guidelines. Version 2. [online] 1998 Oct 27. [cited 1999 Jul 05]. Available from : URL : http://www.ukoln.ac.uk/services/elib/papers/other/standards/

11. Desrichard, Yves. Le Dédoublonnage des banques de données bibliographiques : un état de l'art ». *Documentaliste-Sciences de l'information*, 1997;34(2):82-9.

12. Desrichard, Yves. Les Formats et normes de catalogage : évolutions et perspectives ». *Bull Bibl Fr* 1998; 43(3):56-65.

13. ETC (Electronic Text Center). TEI Guideline for Electronic Text Encoding. [online]. 1994. [cited 1998 Nov 20]. Available from : URL : http://etext.virginia.edu/tei/

14. Gehanno, JF, Paris C, Thirion, Benoît, et al. Assessment of bibliographic databases performance in information retrieval for occupational and environmental toxicology ». *Occup Environ Med*, 1998;55:562-6.

15. Heery, Rachel. Review of Metadata Formats. Program 1996;30(4):345-73. [online].[cited 1998 Nov 20]. Available from : URL : http://www.ukoln.ac.uk/metadata/review.html

16. ICMJE (International Committee of Medical Journal Editors). Uniform Requirements for Manuscripts Submitted to Biomedical Journals. JAMA, 1997;277:927-934. [online] 1997 Mar 18. [cited 1998 Nov 20]. Available from : URL : http://www.ama-assn.org/public/journals/jama/sc6336.htm

17. IFLA (International Federation of Library Associations). UNIMARC: an Introduction. [online]. 1996. [cited 1998 Nov 20]. Available from : URL : http://ifla.inist.fr/VI/3/p1996-1/unimarc.htm

18. ISBD(G). General International Standard Bibliographic Description. Annotated Text. [rev. ed.]. München: K. G. Saur, 1992.

19. ISO 4 : 1997 (revision of 1984). Information and Documentation - Rules for the Abbreviation of Title Words and Titles of Publications. [Geneva] : International Organization for Standardization; 1984.

20. ISO 639 : 1988. – Code for the Representation of Names of Languages. [Geneva] : International Organization for Standardization; 1988.

21. ISO 690 : 1987. Information and

Documentation : Bibliographic References - Content, Form and Structure. [Geneva] : International Organization for Standardization; 1987.

22. ISO 690-2 : 1997. – Information and Documentation - Bibliographic References - Content, Form and Structure. Part 2 : Electronic Documents or Parts Thereof. [Geneva] : International Organization for Standardization; 1997.

23. ISO 2108 : 1992. Information and Documentation - International Standard Book Numbering (ISBN). [Geneva] : International Organization for Standardization; 1992.

24. ISO 2709 : 1996 (revision of 1973). Information and Documentation - Format for Bibliographic Information Interchange on Magnetic Tape. [Geneva] : International Organization for Standardization; 1996.

25. ISO 3166 : 1997 (revision of 1988). Code for the Representation of Names of Countries. [Geneva] : International Organization for Standardization; 1997.

26. ISO 3297 : 1986 (revision of ?). Information and Documentation-International Standard Serial Numbering (ISSN). [Geneva] : International Organization for Standardization; 1986.

27. ISO 8601 : 1988. Data Elements and Interchange Formats - Information Interchange - Representation of Dates and Times. [Geneva] : International Organization for Standardization; 1988.

28. ISO 8879 : 1986 (E). Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML). [Geneva] : International Organization for Standardization; 1986.

29. ISO 9115 : 1987 (cancealed in 1996). Information and Documentation-Bibliographic Identification (BIBLID) of Contributions in Serials and Books. [Geneva] : International Organization for Standardization; 1987.

30. Jolibois, S. ; Nauer, E. ; Mouzé-Amady, M. ; Chouanière, D. ; Grandjean, F. (1999). WebStress Application: a short description of current features and some proposals for future developments. In: Consensus Workshop on Stress at Work, 25-26 Oct 1999, AMI, Copenhagen.

31. Jolibois, S. ; Nauer, E. ; Chouanière, D. ; Mouzé-Amady, M. ; Ducloy, J. ; Grandjean, F. (2000). L'Unified Medical Language System (UMLS): une base de connaissances multilingue dans le domaine biomédical. *Documentaliste-Sciences de l'information*, 2000;37(1).

32. Laisiepen, K. ; Lutterbeck, E. ; Meyer-Uhlenried, K. H. – Grundlagen der praktischen Information und Dokumentation. München : Saur; 1980.

33. Library of Congress (LC). MARC Standards. [online]. Updated 1999 Jan 27. [cited 1999 Feb 23]. Available from : URL : http://lcweb.loc.gov/marc/marc.html

34. Library of Congress (LC). Marc Document Type Definitions. Background and development. [online]. Updated 1998 Mar 2 [cited 1999 Feb 23]. Available from : URL : http://lcweb.loc.gov/marc/marcdtd/marcdtdback.html

35. Lupovici, Catherine. Cataloguer en SGML : de l'étiquetage au balisage. In 63rd IFLA General Conference; 1997 Aug 31-Sept 5. [online]. [cited 1998 Nov 20]. Available from : URL : http://www.ifla.org/ifla/IV/ifla63/63lupcf.htm

36. Lupovici, Catherine. Le Digital Object Identifier : le système du DOI. *Bull Bibl Fr* 1998;43(3):49-54.

37. OCLC. The Dublin Core: A Simple Content Description Model for Electronic Resources. [online]. Updated 1998 May 3. [cited 1999 Feb 23]. Available from : URL : http://purl.oclc.org/dc/

38. OII (Open Information Interchange). Library Information Interchange Standards. [online]. Updated 1998 Aug. [cited 1998 Nov 20]. Available from : URL : http://www2.echo.lu/oii/en/library.html

39. Nauer, E. (1999). De l'importance de la normalisation en bibliométrie. In: 7e colloque sur les systèmes d'information élaborée; 1999 Sep 27th-Oct 1st; Ile Rousse, France.

40. NLM (National Library of Medicine). UMLS Knowledge Sources Documentation. 10th Ed. Bethesda (MD): NLM; 1999. [online]. [updated 1999 Jan 1]. [cited 1999 Jan 29]. Available from : URL : http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML

41. Paskin, Norman. Information Identifiers. Learned Publishing 1997;10(2):135-56. [online] [cited 1998 Nov 20]. Available : from : URL http://www.elsevier.com/inca/homepage/about/infoident/Menu.shtml

42. Patrias, K. National Library of Medicine Recommended Formats for Bibliographic Citation. Bethesda (MD): NLM; 1991.

43. University of Washington. AMA Style Guide [fact sheet online]. 1996 [updated 1997 Jul 8]. [cited 1998 Nov 20]. Available from : URL : http://healthlinks.washington.edu/hsl/styleguides/ama.html

44. University of Washington. Style Guide [fact sheet online]. 1996 [updated 1997 Jul 8]. [cited 1998 Nov 20]. Available from : URL : http://healthlinks.washington.edu/hsl/styleguides/nlm.html