La conception de bases de données infométriques hybrides : analyse de la pratique de trois observatoires européens

Luc Grivel, Hélène Fagherazzi, Philippe Fourneret, Amai Zerouki

La conception de bases de données infométriques hybrides : analyse de la pratique de trois observatoires européens et proposition d'une méthode d'intégration de données hétérogènes

Luc Grivel*, Hélène Fagherazzi**,
Philippe Fourneret**, Alain Zerouki**

* Ingénieur Recherche et Développement à l'Unité Recherche et Innovation
** Ingénieurs documentalistes à la Direction Technique de la Base de Données

Institut de l'Information Scientifique et Technique du Centre National de la Recherche Scientifique (INIST-CNRS)
2, allée du Parc de Brabois
54514 Vandoeuvre-lès-Nancy Cedex.

Tel : (33) 03 83 50 46 00
Fax : (33) 03 83 50 47 33 - Mail : grivel@inist.fr

Mots-clés : infométrie, bibliométrie, base de données, indicateur scientifique et technique, modèle relationnel, modèle objet, observatoire des sciences et techniques, SGML, XML, producteurs de bases de données

Résumé

Les méthodes employées pour le calcul d'indicateurs de politique scientifique sont fondées sur les lois bibliométriques (loi de Zipf pour les mots-clés, loi de Lotka pour les auteurs, loi de Bradford pour les périodiques). Elles s'appliquent en particulier à la littérature scientifique et nécessitent une normalisation des champs de données bibliographiques. Rassemblant des informations scientifiques et techniques normalisées et codifiées, une base est dite 'infométrique' ou 'bibliométrique' lorsque sa structure a été conçue pour obtenir des indicateurs infométriques ou bibliométriques. Il n'existe pas de producteurs directs de bases infométriques mais des bases constituées à partir de données fournies par les producteurs de bases de données bibliographiques.

Le besoin croissant d'indicateurs européens, nationaux, régionaux, institutionnels demande, pour être satisfait, la mise en place de nouvelles bases de données, hybrides (multi-sources), adaptées au calcul d'indicateurs. Comment les concevoir ? Comment les alimenter ?

L'objectif de l'article est double, mettre en évidence quelques points clés et les difficultés pour construire ce type de base et tirer les leçons d'expériences offrant une certaine similarité avec cette problématique.

L'article aborde les problèmes de la couverture et de l'organisation de bases infométriques hybrides en analysant dans un premier temps les pratiques de trois observatoires des sciences et technologies. Après avoir mis en évidence les difficultés liées à l'hétérogénéité des données dans un tel contexte, nous proposons une approche développée dans le cadre de la veille scientifique. Nous en montrons les avantages et les limites pour la constitution de bases infométriques hybrides adaptées au calcul d'indicateurs. Cette approche est basée sur une représentation des documents par une structure d'arbre étiqueté couramment employée pour décrire des documents SGML. La méthode proposée permet de spécifier de manière déclarative les relations entre les éléments de données et leur représentation dans le système de gestion de base de données (SGBD). Cette technique s'intègre parfaitement avec le choix des observatoires de s'appuyer sur les SGBD pour l'exploitation de leurs données. Plus généralement, nous montrons que l'emploi de SGML en association avec un système de gestion de base de données (si possible orienté objet) améliore significativement les possibilités d'exploitation des données. Les autres avantages sont non seulement de permettre l'intégration de données hétérogènes dans une base, mais aussi de distribuer des informations extraites de la base de données sous forme de données SGML pour des traitements ultérieurs ou pour naviguer dans la base infométrique à travers une interface hypertexte.

1 Introduction

On constate depuis quelques années une demande croissante pour des indicateurs permettant de mesurer les activités scientifiques et technologiques, et ce à différents niveaux. Ainsi, selon l'Observatoire des Sciences et Technologie (OST) en France, émergent «*de nouveaux besoins et de nouveaux marchés pour l'infométrie tant au niveau des politiques régionale, nationale, européenne et internationale qu'au niveau du CNRS, des laboratoires, des directions scientifiques, de la direction du CNRS, voire des sections du Comité National* ». Selon son homologue canadien, *« tous les ministères, tant aux États-Unis qu'au Canada (niveau fédéral), doivent proposer des indicateurs de performance dans la description même de leurs programmes. Les programmes et activités relatifs à la science et à la technologie n'échappent pas à la règle. Les universités, au niveau provincial, sont également de plus en plus amenées à produire des indicateurs de résultats.* » (http ://www.ost.qc.ca). En Europe, les instances régionales ont besoin d'outils d'aide à la décision pour déterminer et évaluer leur politique en matière d'innovation, financement de la recherche, etc. Elles jouent en effet un rôle grandissant auprès des acteurs économiques et des acteurs de la recherche par des incitations, par exemple, sous forme de contrats-plans. Au niveau institutionnel, certains organismes (essentiellement des grandes entreprises ou des organismes publics) collectent des données qu'ils souhaitent pouvoir traiter selon des critères infométriques.

Les méthodes employées pour le calcul d'indicateurs de politique scientifique sont fondées sur les lois bibliométriques (loi de Zipf pour les mots-clés, loi de Lotka pour les auteurs, loi de Bradford pour les périodiques). Le calcul d'indicateurs à partir de la littérature scientifique nécessite une normalisation des champs de données bibliographiques sur lesquels s'appliquent les méthodes infométriques. Constatant l'inadéquation des bases de données en ligne pour répondre à ce type de besoins (manque de normalisation, manque d'outils pour les calculs bibliométriques [MOED 1988]), certains observatoires des sciences et technologies ont donc constitué leurs propres bases, dites infométriques, à partir de données fournies par les producteurs de bases de données bibliographiques. Une base infométrique rassemble donc des informations scientifiques et techniques normalisées et codifiées. Sa structure doit être conçue pour faciliter le calcul des indicateurs infométriques ou bibliométriques. Il n'existe pas à l'heure actuelle de producteurs directs de bases infométriques, ni de bases infométriques en ligne.

Le besoin croissant d'indicateurs européens, nationaux, régionaux, institutionnels, que nous avons pu observer à la 5ème conférence internationale des indicateurs scientifiques et techniques Hinxton 1998, demande, pour être satisfait, la mise en place de nouvelles bases de données hybrides (multi-sources), adaptées au calcul d'indicateurs. Comment les concevoir ? Comment les alimenter ?

L'objectif de l'article est double. Mettre en évidence quelques points clés et les difficultés pour construire ce type de base et tirer les leçons sur le plan informatique d'expériences[1] offrant une certaine similarité avec cette problématique. C'est pourquoi cet article comporte deux parties

[1] Par exemple, en 1998, une analyse infométrique de données multi-sources a été mise en œuvre dans le cadre d'une collaboration avec le Bureau Van Dijk (BVD) pour réaliser un rapport de tendance dans le domaine des plantes transgéniques. L'étude a été réalisée sur un corpus de brevets et trois corpus de références bibliographiques issus de PASCAL et d'autres bases de données (AGRICOLA, BIOSIS, EMBASE). Les données ont été stockées dans une base relationnelle par le système HENOCH. [POLANCO 98]

distinctes. La première ne nécessite pratiquement aucune connaissance en informatique et peut se lire indépendamment de la deuxième. A l'inverse, la deuxième s'adresse plutôt à des informaticiens mais requiert la lecture de la première partie pour comprendre le contexte d'application. La première partie (section 2) décrit la couverture et de l'organisation générale des bases infométriques en se basant sur les pratiques d'observatoires des sciences et technologies dans trois pays européens (la Hollande, la France et l'Espagne). Il ne s'agit pas de comparer ces trois observatoires mais de décrire ce qui caractérise une base infométrique de nos jours. Les problèmes relatifs à la constitution de tels bases sont mis en évidence. L'un de ces problèmes, l'hétérogénéité des données, constitue le sujet d'étude de la deuxième partie (section 3). Il y est décrit une méthode d'intégration de données hétérogènes développée dans un contexte de veille scientifique. Cette méthode utilise des techniques informatiques de gestion documentaire. Nous en montrons les avantages et les limites pour la constitution de bases infométriques hybrides adaptées au calcul d'indicateurs.

2 Bases de données infométriques

Nous avons choisi comme source d'exemples trois observatoires européens, représentatifs sur le plan international, qui ont décrit leur base infométrique dans des publications scientifiques : un pays largement anglophone : la Hollande, et deux pays de langue latine, l'Espagne et la France. Un tableau descriptif des observatoires sur le plan des missions, ressources, indicateurs produits figure en annexe II.

2.1 Présentation des organismes et de leurs objectifs

a) L'Espagne

L'Espagne dispose avec le CINDOC, centre de documentation scientifique du CSIC, (Consejo Superior de Investigaciones Cientifica, http://www.cindoc.csic.es) d'un organisme comparable à l'INIST en France. Parmi ses missions figure la réalisation d'études bibliométriques en tant qu'outils d'aide à la définition d'une politique scientifique et à l'évaluation des programmes scientifiques espagnols [FERNANDEZ 93, BORDONS 95, GOMEZ 95].

b) La France

La France a créé en 1990 l'Observatoire des Sciences et Technologie (OST), groupement d'intérêt public chargé de fournir des éléments d'analyse sur les activités de recherche et de développement technologique en France. L'OST a construit sa propre base de données infométriques avec comme objectif « la construction d'indicateurs fiables, pertinents et pérennes, décrivant la science et la technologie française » en comparaison européenne ou internationale [BARRE 95, Rapport OST 1998, ZITT 1996].

c) La Hollande

La Hollande a créé en 1992 le NWOT (Netherlands Observatory of Science and Technology) qui coordonne la collaboration de deux équipes pour la publication du Netherlands S&T Indicators Report : le CWTS (Centre for Science and Technology Studies(http://sahara.fsw.leidenuniv.nl/) et le MERIT (Maastricht Economic Research Institute on Innovation and Technology). Leur rapport 1998 est disponible sur Internet (http://sahara.fsw.leidenuniv.nl/cwts/summary.html).

Centre de recherche dans le domaine de l'analyse quantitative de la recherche, le CWTS est à l'origine de la conception de la base infométrique permettant l'élaboration et l'application d'indicateurs dans le domaine de la recherche scientifique et technologique aux Pays-Bas. [MOED 1988, 1995, 1996].

2.2 Données et structure de données dans les bases infométriques

Nous mettrons l'accent dans cette sous-section sur ce qui caractérise une base infométrique. Les méthodes pour réaliser des indicateurs à partir de données bibliographiques vont de la statistique descriptive aux analyses multidimensionnelles, en passant par des techniques de classification et de cartographie ; [ROSTAING 96] constitue une bonne introduction à ces méthodes. L'aspect calcul et type d'indicateurs est abordé plus complètement dans [MOED 96], [GLANZEL 96].
On peut observer que la plupart des indicateurs publiés dans les rapports des trois observatoires étudiés sont des indicateurs univariés[2]. Les indicateurs relationnels les plus couramment utilisés sont les co-publications et cocitations, en se limitant à du dénombrement Les indicateurs les plus sophistiqués (classification, cartographie) ne sont employés que dans le cadre d'études à la demande (voir annexe II).

2.2.1 Données

Le plus souvent, les études infométriques qui sont menées par ces observatoires utilisent une source de référence unique (les bases de l'ISI). L'ISI fournit aux observatoires un fichier, l'Integrated Citation File, (ICF) qui est une compilation structurée de ses différentes bases (SCI, SSCI, A&HCI, voir en annexe I, un exemple de fiche bibliographique extraite du SCI.). La caractéristique de l'ICF est de constituer une base où documents citants et documents cités sont appariés, formant un réseau de documents se citant les uns les autres.

Pour donner un exemple sur la manière de procéder, voici comment est constituée la base infométrique de la Hollande. L'ISI a fourni toutes les publications du SCI, SSCI, A&HCI à partir de l'année 1980 à 1993 comportant des adresses d'auteurs originaires de Hollande. Dans chaque publication figurent tous les auteurs de la publication, leurs adresses, les données sur la source (titre du périodique, année, numéro de volume, pagination, type de document), le titre de la publication, les références citées. Sont fournies également toutes les publications issues des mêmes bases citant ces publications hollandaises pendant la même période. La base est ensuite mise à jour tous les deux ans.

L'OST utilise une version simplifiée de l'Integrated Citation File qui signale pour chaque publication les éléments catalographiques (journal, date de publication, ...) et surtout les pays d'origine de l'article tels qu'ils sont repérées dans les adresses d'auteur, complétées pour les adresses européennes par les codes postaux, le nombre de citations reçues sur les 2 et 5 années suivantes, par pays citant.

[2] Chaque élément à étudier est soumis à une mesure selon une dimension choisie (dénombrement, calcul de ratio)

Pourquoi les observatoires procèdent ils de cette manière ?

Se plaçant sur le plan de la production d'indicateurs, les observatoires cherchent à développer des bases infométriques répondant à deux critères principaux du point de vue de leur couverture:

- une couverture très sélective au niveau des périodiques (revues cœur) et stable dans le temps ;
- une couverture multidisciplinaire pour pouvoir comparer les disciplines ou domaines et couvrir des thématiques pointues.

Une telle couverture permet des comparaisons dans le temps, en garantissant que le choix de revues répond à des critères qualitatifs clairs et contrôlables (facteur d'impact, comités d'experts, etc.).

Actuellement le SCI est la seule base multidisciplinaire répondant globalement à ces critères. Le Science Citation Index de l'ISI est donc la source par excellence pour les études infométriques à partir des publications scientifiques.

Les qualités qui ont fait du SCI la base de référence sont d'après [BARRE 95, Rapport européen 97] :

- multi-disciplinarité (tous les domaines de recherche y sont bien représentés, à part les sciences sociales et les mathématiques, couvertes respectivement par le SSCI et CompuMath, produites également par l'ISI)
- sélectivité (sélection des périodiques selon une mesure d'impact et selon avis d'un comité d'experts)
- traitement complet des périodiques (cover to cover) : tous les documents issus du périodique sont enregistrés dans la base, qu'il s'agisse d'articles 'normaux', de synthèses, de notes, de lettres, etc.
- en principe, complétude des auteurs et des adresses (utilisées pour l'analyse des collaborations scientifiques)
- citations (toutes les références bibliographiques sont saisies, permettant une analyse des citations)
- disponibilité dans un format exploitable infométriquement (l'Integrated Citation File).

Ses principaux défauts [Rapport européen 97][DOUSSET 97] sont :

- couverture inégale ou discutable de certains domaines scientifiques (sciences appliquées, notamment les sciences pour l'ingénieur ou la pédologie), et déséquilibre entre les disciplines (sur-représentation de la médecine clinique par exemple).
- origine essentiellement anglophone des publications qu'elle signale,
- forte coloration américaine, ce qui implique que la recherche européenne ne s'y trouve que partiellement représentée,
- absence de normalisation des auteurs citants et cités et des titres des revues. Ces données saisies à l'état brut doivent faire l'objet de nombreuses corrections.
- pas d'indexation au niveau article. Cet aspect est en partie compensé par les mots-clés d'auteurs, lorsqu'ils sont présents, et les mot-clés rassemblés sous le champ keywords+

(Indexation automatique sur les titres des articles cités et les notes de bas de page des auteurs).

Les autres bases bibliographiques, quelles soient spécialisées (INSPEC pour la physique, l'électronique et informatique, CAB pour la chimie, MEDLINE pour la médecine, etc.) ou multidisciplinaire (PASCAL), bien que signalées comme étant utilisées par le CINDOC et l'OST, ne sont en fait employées que marginalement. Ces bases sont sous utilisées du point de vue exploitation infométrique.

Les points les plus critiques sont selon les observatoires et dans cet ordre :
- une absence de politique claire concernant la couverture
- la saisie incomplète des auteurs,
- l'absence des citations.

Bien entendu, ces points faibles sont variables selon les bases. Des bases comme MEDLINE ou INSPEC sont reconnues disposer d'une couverture satisfaisante dans leur domaine. PASCAL saisit depuis 1996 les adresses de tous les auteurs. En l'état, les bases de l'INIST offrent donc déjà un certain nombre de caractéristiques intéressantes pour l'analyse bibliométrique, notamment pour les observatoires européens (multi-disciplinarité, indexation par des mots-clés, complétude des adresses des auteurs, couverture plus européenne que le SCI) mais souffrent de l'absence des citations et surtout du manque de clarté concernant la définition de sa politique de couverture. Sur le plan de la littérature cœur, le recouvrement entre les deux bases n'est pas encore tout à fait satisfaisant et des progrès restent à faire.
Concernant le dernier point, les citations sont bien sûr indispensables pour le calcul d'indicateurs d'impact et notamment le facteur d'impact : nombre moyen de citations dont les publications d'une revue font l'objet. Mais dans la pratique, les indicateurs de productivité des chercheurs, des équipes, des institutions ou pays sont les plus simples mais aussi les plus importants des indicateurs [VINKLER 96].

2.2.2 Tables de nomenclatures / fichiers d'autorité

Rôle des fichier d'autorité : agréger et normaliser

Les fichiers d'autorité ou tables de nomenclatures sont indispensables pour définir les niveaux d'agrégation pour les comptages (données numériques) permettant de construire les indicateurs selon des critères géographiques (pays, régions), thématiques (disciplines scientifiques SCI, domaines technologiques) ou selon les secteurs d'activité industrielle.

Ces fichiers jouent également un rôle utile dans la nécessaire phase de normalisation des données bibliographiques avant leur stockage dans la base. Les mêmes données se présentant souvent sous différentes formes lexicographiques, les fichiers d'autorité permettent l'établissement de listes de correspondance, par exemple, pour les noms de pays. La technique généralement utilisée pour établir des équivalences et uniformiser les champs de données présentant des variations essentiellement typographiques (majuscule, minuscule, etc.) ou flexionnelles (pluriels, singuliers) est d'aboutir à une convergence par rapport à une forme appauvrie, analogue à une clé à laquelle est associée sa forme attestée.

Quelques exemples de fichiers d'autorité ou tables de nomenclatures

Disciplines/ domaines scientifiques

La plupart des indicateurs publiés dans les rapports des trois organismes s'appuient sur la classification en discipline de l'ISI. Cette classification définit des catégories « journal categories » où sont regroupés des périodiques qui suivent une spécialité, en anglais, « subfield » (par exemple, optique, botanique, etc.) qui peuvent former ensuite des disciplines « field » (physique, sciences de l'univers, sciences pour l'ingénieur, etc.). L'inconvénient majeur de cette approche est que le groupe de périodiques appartenant à une catégorie particulière peut varier d'une année à l'autre. En outre une classification au niveau d'un périodique, qui est ensuite répercutée à tous les articles de ce périodique, ne peut être aussi pertinente qu'une classification effectuée article par article. L'avantage est que les études utilisant cette nomenclature sont comparables. La classification de l'ISI est de fait devenue une sorte de classification pivot avec d'autres systèmes de classification. L'OST par exemple a construit sa propre classification en 8 disciplines à partir de la classification de l'ISI.

Les indicateurs basés sur des classifications thématiques au niveau 'article' sont plus rarement utilisés même si on leur reconnaît de nombreuses qualités intrinsèques (souplesse dans la définition du domaine, pertinence, etc.). Leur emploi est réservé aux études effectuées sur des données issues de bases qui 'indexent' au niveau article. C'est le cas de la plupart des bases de données spécialisées (INSPEC pour la physique, CAB pour la chimie, MEDLINE pour la médecine, etc.) et de la base multidisciplinaire PASCAL.

Entité géographique/institutionnelle

Dans la plupart des indicateurs, l'unité d'analyse (l'objet d'étude) est une entité géographique ou institutionnelle. Les publications sont assignées à ces unités sur la base d'une analyse des adresses des auteurs. Au sein de données bibliographiques, les variations de noms de pays sont limitées en nombre. Comme le souligne [MOED 96], mettre en correspondance publications et institutions de recherche est une tâche beaucoup plus délicate qui ne peut être effectuée directement et simplement en se basant sur les adresses des auteurs des publications. Très fréquemment, il arrive de rencontrer de nombreuses formes lexicographiques pour la même donnée.

Ceci suppose l'existence de fichiers d'autorité géographiques (codes postaux, villes, régions, pays) et institutionnels (code d'institution, classification sectorielle des organismes, ...).

Chaque organisme s'est donc doté de fichiers d'autorité :

Espagne
Pour le traitement des affiliations, le CINDOC a constitué les fichiers d'autorité suivants :
I/-Centres de recherche
- Nom standardisé
- Code institution
  pour les centres espagnols à 5 niveaux :
    1. *dépendance administrative*

2. *type d'organisation* à l'intérieur de chaque dépendance administrative. (Un code pays en trois lettres est introduit ici pour les centres étrangers)
3. *acronyme*
4. *code UNESCO* disciplinaire
5. *code postal*

NB : les centres étrangers sont codifiés à un niveau plus agrégé

II/-Villes espagnoles (variations des noms, et code postal indiquant la province et la communauté autonome)
III/-Pays étrangers (codes pays anglais et espagnols, code ISO, avec agrégations pour les pays du royaume uni ou les deux anciennes Allemagnes, ainsi que pour des régions multinationales telles que l'Union Européenne et l'Amérique latine)

France
L'OST effectue des regroupements géographiques à divers niveaux d'agrégation (monde, continent, zones du monde, pays, régions (françaises et européennes) en utilisant les adresses postales. L'OST ne constitue pas de fichiers d'autorité concernant les laboratoires de recherche, considérant que cet acte n'est pas de sa responsabilité.

Hollande
Pour résoudre le problème de variation des noms des instituts de recherche hollandais, le CWTS constitue un fichier d'autorité rassemblant pour chaque institution les différentes variations sous une dénomination commune. Cette opération est particulièrement lourde car pour éviter toute controverse, le CWTS compare les adresses apparaissant dans le SCI et celles figurant dans différents répertoires (répertoire des universités, répertoire des organisations de recherche, etc.) et enfin consulte les spécialistes dans les différents domaines de recherche pour valider les résultats obtenus.

Le CWTS a également constitué un système de classification des organismes de recherche néerlandais en trois secteurs :

- public (universités, instituts de recherche, etc...)
- privé (entreprises, etc...)
- « intermédiaire » (pharmacies, etc...)

Facteur d'impact du périodique

Le Journal Citation Reports (JCR) propose le classement d'un ensemble de périodiques scientifiques selon plusieurs critères :

- par domaines (désignés par l'ISI)
- par fréquence de citations : nombre de fois où sont cités les articles publiés par un périodique
- par facteur d'impact : nombre moyen de citations dont les publications d'une revue font l'objet.

Le JCR est de moins en moins utilisé. Les trois organismes recalculent le plus souvent leur propres indicateurs d'impact à partir de l'ICF Integrated Citation File [SMALL 95], certaines

études ayant montré que les facteurs d'impacts publiés par le JCR ne sont pas exacts pour certains périodiques [MOED 95b].
En outre, il existe différentes méthodes pour calculer le taux de citation attendu d'une unité d'analyse (au sens défini plus haut), en anglais, *expected citation rate*, selon qu'il est pondéré ou non par le nombre d'articles publiés par cette unité dans chaque périodique.
Exemple extrait de [MOED 96], supposons que l'unité A ait publié 5 articles dans deux périodiques P1 et P2, 1 dans P1, 4 dans P2 et que le taux moyen de citation (le facteur d'impact) soit respectivement de 4.00 pour P1 et de 9.00 pour P2.
Alors le taux de citation attendu pour l'unité A sera de 8.00 s'il est pondéré par le nombre d'articles et de 6.5 s'il ne l'est pas.

2.3 Modélisation et stockage des données infométriques

Les observatoires désirent analyser tout élément de données ou combinaison d'éléments (auteur, titre, source, affiliation, pays, mots-clés, année de publication, etc.). Comme les bases de données relationnelles ont été conçues explicitement pour relier des éléments de données, elles sont un choix naturel pour les analyses bibliométriques. Technologie éprouvée datant des années 70, leur emploi en infométrie est relativement récent (début des années 90). Les principes de bases du modèle relationnel sont :

- représentation des données sous forme de tables,
- manipulation de ces données à l'aide d'opérateurs appliqués aux tables pour fournir d'autres tables dans le cadre d'une algèbre relationnelle (langage SQL)

L'intérêt majeur d'une telle structuration relationnelle est que les informations provenant de tables présentant un champ commun (numéro d'article, auteur, pays, titre de journal) quelles proviennent ou non d'une même source, sont potentiellement combinables. Ainsi la plupart des indicateurs à produire peuvent être calculés par de simples commandes SQL. Une requête telle que « compter le nombre de documents produits par chaque pays d'affiliation des auteurs et trier les pays par fréquence décroissante » s'écrit facilement en SQL. Le lecteur intéressé trouvera dans [BLAIR 88] de nombreux exemples de requêtes de ce type implémentées en SQL. Des tables réceptionnent les résultats des opérations de croisement nécessaires pour le calcul des indicateurs.

Chaque élément d'information (titre de périodique, auteur, etc.) de chaque document alimente la table lui correspondant (table des périodiques, table des auteurs, etc.).

Chaque document est identifié par une clé (NuméroDocument), c'est à dire un numéro, attribut qui le relie aux auteurs, aux institutions et au journal où l'article a été publié.
Les fichiers de nomenclatures sont également mis sous forme de tables, comme par exemple la classification des périodiques par catégorie.

Les trois observatoires stockent leurs données dans une base relationnelle afin de réaliser, par des requêtes SQL, les croisements à effectuer pour calculer les indicateurs. Les volumes de données stockés sont de l'ordre de plusieurs millions de documents.

2.4 Conclusion

Nous venons de décrire les données et structures de données qui caractérisent les bases infométriques de trois observatoires (fichiers d'autorité, données bibliographiques normalisées, modélisation relationnelle) en explicitant les raisons de leurs différents choix.

Sur le plan méthodologique, les points clés sont :

1. une couverture multi-disciplinaire, très sélective, à l'instar de ce que fait l'ISI au niveau des périodiques (revues cœur), et stable dans le temps, tout en garantissant une bonne représentativité des différents domaines. La couverture optimale d'une thématique nécessite une démarche multidisciplinaire. Ce qui suppose un élargissement des domaines couverts. Cette couverture doit être évaluée périodiquement (facteur d'impact, comité d'experts, indicateurs infométriques, etc.)
2. la constitution et l'utilisation de tables de nomenclatures pour réaliser divers indicateurs selon des critères géographiques (pays, régions) ou thématiques (disciplines scientifiques, domaines technologiques) ou selon les secteurs d'activité industrielle,
3. la structuration et la normalisation de différents champs de données (journaux, adresse d'affiliation des auteurs, noms des auteurs, ...) en s'appuyant sur des fichiers d'autorité et/ou des règles de normalisation,
4. une modélisation des données adaptée au calcul d'indicateurs.

Dans le contexte des observatoires, les volumes de données stockés sont de l'ordre de plusieurs millions de documents. Les trois observatoires stockent leurs données dans une base relationnelle afin de réaliser, par des requêtes SQL, les croisements à effectuer pour calculer les indicateurs.

A notre connaissance, si on en juge par les études effectuées, il n'y a pas réellement intégration de données hétérogènes dans un modèle de données commun. Les données proviennent généralement d'une même source (l'ISI). Si une étude requiert exceptionnellement des données provenant d'autres sources, elles sont traitées et stockées séparément des données de l'ISI. Pourtant, les observatoires étudiés reconnaissent implicitement qu'un élargissement des sources utilisées leur permettrait de répondre de manière plus satisfaisante aux multiples niveaux de demande. Quels sont les obstacles à la construction de bases infométriques hybrides (multi-sources) ?

Ils sont à la fois techniques et juridiques. Sur le plan technique, une base infométrique hybride suppose une véritable intégration des données dans le SGBD. On se rapproche ici des problématiques de la gestion de bases documentaires où le besoin de transformer les documents pour pouvoir les partager entre applications a toujours été une préoccupation majeure. Les apports de ces techniques sont développés dans la section suivante où nous abordons la question de l'hétérogénéité des données et des formats, et donc de la normalisation. Nous abordons également la question de la modélisation des données et de l'environnement informatique.

Les autres obstacles sont de nature plus politique ou juridique. Par exemple, pour définir une couverture élargie, il est nécessaire d'interroger plusieurs bases de données. Certains producteurs de données refusent ou font payer très cher la constitution de nouvelles bases à partir de données leur appartenant, imposant une licence à un coût élevé et/ou se donnant un

droit de regard sur l'utilisation de ses données. Autre exemple : la constitution de fichiers d'autorités pour les organismes d'affiliation. Sans la collaboration des organismes concernés, il est difficile d'établir des fichiers pertinents. La fourniture d'un organigramme simplifie la tâche, de la même manière qu'il est plus facile de faire une normalisation des descripteurs (mots-clés) si on dispose de ressources terminologiques[3].

A travers ce constat, se pose le problème de la définition des relations producteur de bases de données - observatoires et producteurs de bases de données entre eux, sans oublier les auteurs/organismes qui sont à l'origine des publications. Sans compétence particulière sur le plan juridique, nos réflexions se limitent à exprimer une opinion. Construire des bases infométriques hybrides ne peut s'envisager sans mettre en place un cadre de coopération équitable entre les producteurs de bases de données et les observatoires, les instituts de recherche pour définir la couverture des bases, améliorer la normalisation des données, constituer ou utiliser des fichiers d'autorités communs en partageant coûts, compétences et forces de travail.

[3] Sur ce dernier point, signalons les travaux de J. Royauté sur les groupes nominaux complexes [ROYAUTE 99] et leurs propriétés, et notamment son étude du phénomène de la variation en corpus, quelles soient flexionnelles ou syntaxiques. Ces travaux ont débouché sur une plate-forme linguistique (ILC) qui permet de repérer des termes en corpus sous leurs différentes formes en liaison avec un lexique terminologique.

3 Intégration de données héterogènes

L'objectif de cette deuxième partie est de tirer les leçons de diverses expériences de veille[4] que nous avons menées. L'URI a développé une approche originale basée sur un couplage SGML/SGBD qui permet de construire et d'exploiter des indicateurs infométriques dans un environnement hypertexte convivial à des fins de veille scientifique, en employant une méthodologie un peu analogue à celle des observatoires des sciences et techniques (section 2) et des méthodes de traitement de données issues du monde de la gestion documentaire. Ces travaux ont débouché sur une plate-forme infométrique dont l'un des composants, le logiciel HENOCH, permet d'intégrer des données hétérogènes en types et en formats [GRIVEL 95,97,99], cf annexe 3).

Ces expériences ont nécessité l'intégration de données hétérogènes dans une base de données relationnelle qui est, comme nous l'avons vu une des difficultés de la construction de bases infométriques hybrides.

Alimenter un SGBD à partir de documents fait partie des applications courantes dans le monde documentaire. D'une manière générale, il s'agit de transformer un document d'une certaine structure logique en une autre. L'intérêt de SGML/XML[5] dans ce contexte n'est plus à démontrer. On trouve aujourd'hui sur le marché plusieurs éditeurs SGML/XML disposant d'une interface avec les principaux SGBD du marché [MICHARD 98]. Il est ainsi possible, en utilisant les interfaces de programmation (API) de l'éditeur SGML/XML et du SGBD, de développer une passerelle de stockage dans la base de donnée de tout élément XML 'parsé' (analysé) par l'éditeur.

L'approche la plus commune, couramment utilisée par la plupart des parseurs (analyseurs) de documents SGML, est d'extraire la structure des documents en passant par un modèle pivot intermédiaire, le plus souvent, une structure d'arbre étiqueté. La totalité du document est alors représentée dans cette structure d'arbre étiqueté.

L'approche que nous exposons ici s'inspire de cette méthode. Elle est de prendre les documents dans leur structure logique initiale, traduite le plus fidèlement possible dans le format SGML, en extrayant les données qui nous intéressent dans un SGBD relationnel selon une méthode qui permette de tenir compte à la fois des données représentées dans une structure d'arbre et des données existant dans la base.

---

[4] Par exemple, en 1998, une analyse infométrique de données multi-sources a été mise en œuvre dans le cadre d'une collaboration avec le Bureau Van Dijk (BVD) pour réaliser un rapport de tendance dans le domaine des plantes transgéniques. L'étude a été réalisée sur un corpus de brevets et trois corpus de références bibliographiques issus de PASCAL et d'autres bases de données (AGRICOLA, BIOSIS, EMBASE). Les données ont été stockées dans une base relationnelle par le système HENOCH. [POLANCO 98]

[5] SGML, Standard Generalised Mark Up Language, norme [ISO 8879], [GOLDFARB 90], HERWIJNEN 90], Le format SGML (Standard Generalized Markup Language) donne des règles de balisage pour décrire des structures arborescentes où chaque noeud est identifié par une étiquette. Baliser un document consiste à insérer dans le texte des chaînes de caractères qui donnent de l'information sur le contenu du document.
XML (eXtensible Markup Language) est une version modernisée et simplifiée de SGML, issue des travaux du W3C. XML retient les caractéristiques essentielles de SGML en l'épurant de ses caractéristiques les plus complexes à mettre en œuvre et en apportant de puissants de mécanismes de liens, étendant ceux présents dans HTML. Il existe une traduction en français de la norme XML, http://babel.alis.com/web_ml/xml

Peut on facilement transposer cette approche développée dans un contexte de veille à l'échelle des bases infométriques des observatoires des sciences et techniques ?

Nous exposons ici notre méthode et nous l'évaluons.

3.2 Structure de données, normalisation et modèle de données : une approche intégrée pour résoudre les problèmes d'hétérogénéité des données et des formats

3.2.1 Reformatage

Dans le cas de notices bibliographiques, la sémantique est exprimée dans les étiquettes décrivant les champs, et éventuellement par l'ordre des données. En utilisant un analyseur lexical, on peut aisément décrire au format SGML/XML des notices bibliographiques déchargées à partir d'un serveur de données, sans perdre d'informations [DUCLOY 91]. La structure logique d'une notice bibliographique telle que celle décrite en annexe 1, est très simple : une suite de champs repérés par un identifieur. Il est relativement facile de définir les règles lexicales qui permettent d'identifier le début ou la fin d'une notice, le début ou la fin d'un champ à l'intérieur de la notice de manière à la transformer en document SGML en forme normale.

<record>
<NO>12508319 </NO>
<TI>AMYOTROPHIC-LATERAL-SCLEROSIS AND STRUCTURAL DEFECTS IN CU,ZN SUPEROXIDE-DISMUTASE </TI>
<AU> DENG HX; HENTATI A; TAINER JA; IQBAL Z; CAYABYAB A; HUNG WY; GETZOFF ED; HU P; HERZFELDT B; ROOS RP; WARNER C; DENG G; SORIANO E; SMYTH C; PARGE HE; AHMED A; ROSES AD; HALLEWELL RA; PERICAKVANCE MA; SIDDIQUE T
</AU>
<AF><NA> NORTHWESTERN UNIV,SCH MED,DEPT NEUROL,300 E SUPER ST NEUROL</NA><TO>CHICAGO</TO><CO>IL</CO></AF> ...
</record>

3.2.2 Intégration des données dans un SGBD : méthode

Une fois les données reformatées, il faut ensuite les intégrer dans un modèle de données. En s'appuyant sur la structure d'arbre des documents SGML, il est possible de définir la correspondance entre les attribut de chaque table constituant la base relationnelle et des chemins d'accès aux éléments de données et d'associer un traitement particulier à ces données : une procédure qui réalise les tests et actions nécessaires pour interpréter la chaîne de caractère correspondant à l'élément de données en fonction du modèle de données de la base

La structure d'arbre permet un accès direct à tout noeud de l'arbre. Nous avons défini une sorte de grammaire annotée qui permet d'associer une variable à un noeud, cette variable étant un paramètre d'une procédure (PL/SQL en l'occurrence), qui est exécutée lorsque tous ses paramètres sont instanciés. Un noeud (élément de données dans la terminologie SGML) peut

être qualifié par un symbole d'occurrence. Par exemple, un noeud déclenche autant d'appels de la procédure qu'il y a de valeurs répétitives (c'est le cas par exemple d'une liste de mots-clés ou d'affiliations).

Un fichier de configuration associé à un type de document décrit la mise en correspondance entre les variables et les différents champs de la notice.

Dans l'exemple ci-dessous, ce fichier décrit comment alimenter une table des affiliations à partir d'un document reformaté comme celui de la section 3.1.1 :

Nom de la variable	**Chemin d'accès à un noeud de l'arbre**	**occurrence**
Name	record/AF/NA	repeat
Town	record/AF/TO	repeat
Country	record/AF/CO	repeat

query :
begin
/* the insertion procedure to execute */
INS_AFFILIATION(:{NAME}, :{TOWN}, :{COUNTRY})
end;

Avant de stocker les informations dans la base, la procédure d'insertion effectue les tests nécessaires pour, par exemple, vérifier si le nom du pays est bien conforme à un nom de pays figurant dans la table des noms de pays, tenter d'apparier la chaîne de caractère représentant le nom de l'organisme avec la table des noms d'organismes, etc.

Cette approche spécifie donc de manière déclarative les relations entre les éléments de données et leur représentation dans la base en utilisant une sorte de 'règle de réécriture' qui permet d'exécuter, par exemple une méthode de création d'un objet complexe (par exemple une super-notice[6] bibliographique) à partir des éléments de données.

3.3 Evaluation

Ce procédé a été implanté dans le logiciel HENOCH [GRIVEL 95, 97, 99] dans un contexte de veille où le nombre de documents à gérer ne dépasse pas quelques milliers de documents.

Cette méthode est plus efficace qu'une interprétation directe du fichier de données qui se contenterait de stocker l'élément de données sous forme de chaîne de caractères (string) directement dans la base. Elle permet d'éviter la présence d'informations inutiles dans cette chaîne de caractère en la traitant avant de la stocker dans la base, et de pallier à l'absence

[6] Dans le cas de données multi-sources, la présence de doublons est inévitable. Au lieu d'éliminer les doublons en ne gardant qu'un exemplaire de notice pour chaque clé, en privilégiant par exemple un ordre de préférence dépendant de la base d'origine [NAUER 99], les doublons peuvent être utilisés pour construire des « super-notices », en prenant par exemple, tel champ d'une source et tel autre d'une autre source, ou en combinant deux champs, sur la base de la présence ou de l'absence de telle ou telle information (cf annexe 3)

d'information dans la chaîne elle-même, en allant, si nécessaire, chercher de l'information dans d'autres éléments de données, des index ou dans la base.

La technologie utilisée dans HENOCH au niveau de la procédure d'insertion, une procédure écrite en PL-SQL, a un inconvénient principal : dans la phase de stockage, elle effectue des tests sur le contenu de chaînes de caractères stockées dans le SGBD. Elle utilise les méthodes de recherches du SGBD qui sont moins performantes que les systèmes basés sur les index.

Cette limite est inhérente à la technologie de la plupart des SGBD relationnels : ils n'indexent pas les structures de données de type *string*. Lorsque nous avions développé HENOCH, nous ne nous étions pas posés le problème en ces termes. L'idée était simplement de pouvoir stocker facilement quelques milliers de documents issus de différentes sources au format SGML ainsi que les résultats de classifications sur ces données. Dans le contexte des observatoires, une solution plus efficace consisterait à coupler un moteur d'indexation et de recherche au système de gestion de bases de données.

Sur de très gros volumes de données (ce qui est le cas des bases infométriques des observatoires), un couplage XML-SGBD Orienté Objet serait, sans doute, mieux adapté qu'un couplage XML-SGBD relationnel. En effet, dans le modèle relationnel, la représentation plate d'un document structuré tel qu'une notice bibliographique se paie par un coût qui peut vite devenir rédhibitoire pour de grands volume de données. Lorsqu'il s'agit de 'reconstruire' une notice à partir de ses éléments, le modèle objet est plus efficace puisqu'il permet de représenter directement la hiérarchie des éléments et l'héritage des propriétés dans l'arbre représentant le document [MICHARD 98]. En effet, dans le modèle objet, on dispose de deux mécanismes d'accès à un objet [DUCOURNEAU 98] : un mécanisme d'accès par contenu comme dans un SGBD relationnel et un mécanisme d'accès par référence utilisant ses liaisons logiques avec d'autres objets. Chaque fois qu'un nouvel objet (par exemple, un élément de la notice) est créé dans la base, il est possible de lui donner un identificateur et de le retrouver directement dans une transaction. Les identificateurs des objets avec lesquels un objet O est en relation par héritage permettent au système d'assurer à moindre coût la recomposition de l'objet en utilisant les liaisons de O.

La technique proposée devrait donc être plus efficace dans un environnement couplant XML, un moteur d'indexation et de recherche d'information et un SGBDOO.

D'un point de vue pragmatique, le couplage XML et SGBD, que ce dernier soit relationnel ou objet, est, *de toute façon*, une solution qui permet de bénéficier du meilleur de ces deux technologies. Elle permet non seulement l'intégration de données hétérogènes dans une base, mais aussi de distribuer des informations extraites de la base de données sous forme de données XML, soit pour des traitements ultérieurs, soit pour naviguer dans la base infométrique à travers une interface hypertexte. Elle est viable sur le long terme, d'autant plus que chacun des deux types d'environnement propose des interfaces de programmation (API) qui tendent à se standardiser.

4 Conclusion

L'un des problèmes relatifs à la constitution de bases infométriques est l'hétérogénéité des données. Nous avons proposé une approche informatique basée sur un couplage XML/SGBD pour l'intégration de données hétérogènes. Cette approche spécifie de manière déclarative les relations entre les éléments de données et leur représentation dans la base en utilisant une sorte de 'règle de réécriture' qui permet d'exécuter, par exemple une méthode de création d'un objet complexe à partir des éléments de données.

Nous avons en montré les avantages et les limites pour la constitution de bases infométriques hybrides adaptées au calcul d'indicateurs. La technique proposée permet d'éviter la présence d'informations inutiles dans la base, et de pallier à l'absence d'information dans la chaîne elle-même, en allant, si nécessaire, chercher de l'information dans d'autres éléments de données, des index ou dans la base. Cette technique, testée dans un environnement SGML/SGBD relationnel serait plus efficace dans un environnement couplant SGML, un moteur d'indexation et de recherche d'information et un SGBDOO.

D'une manière générale, l'emploi de SGML/XML en association avec un système de gestion de base de données (si possible orienté objet) améliore significativement les possibilités de d'exploitation des bases données documentaires existantes (bibliographiques, brevets, etc.), ce qui devrait permettre de répondre plus complètement aux multiples niveaux de demande.

Nous avons appris récemment qu'un procédé, similaire dans l'esprit à celui que nous avions mis en place dans le système HENOCH mais basé sur la technologie objet, était mis en oeuvre pour charger des données hétérogènes dans un SGBDOO, O2 [ABITBOUL 97]. Ce n'est pas trop surprenant. L'intégration de données hétérogènes au sein d'un SGBD est un champ de recherche très actif dont le champ d'application a pris une surface considérable avec l'essor du Web. Ce champ de recherche n'a pas réellement retenu l'attention des infométriciens dont la préoccupation première est de définir de nouvelles méthodes de calculs d'indicateurs. Pourtant la fiabilité de ces calculs repose en partie sur la capacité à résoudre les problèmes liés à l'hétérogénéité des données. Il est donc important de s'appuyer sur les techniques les plus avancées des systèmes de gestion de bases de données.

BIBLIOGRAPHIE

[ABITEBOUL 97] Querying Documents in Object Databases, Serge Abiteboul, Sophie Cluet, Vassilis Christophides, Tova Milo, Guido Moerkotte, Jerome Simeon, International Journal on Digital Libraries, 1(1), 5-19, 1997.

[BARRE 95] BARRE R., LAVILLE F., TEIXEIRA N., ZITT M. 'L'observatoire des sciences et des techniques : activités- définition- méthodologie' SOLARIS, 1995, 2, p.219-235.

[BLAIR 88] BLAIR D.C. 'An extended relational Document Retrieval Model', Information Processing and Management Vol 24, n°3 (1988), 259-371.

[BORDONS 95] BORDONS M. ., ZULUETA M.A, CABRERO A . 'Identifying Research teams with bibliometric tools publications' In: Michael E.D Koenig, Abraham Bookstein (Eds), 5th International Conference of the International Society for Scientometrics and Informetrics, Learned Information Inc. Medford NJ, 83-92.

[DOUSSET 97] DOUSSET B., DKAKI T. 'Evaluation et expertise scientifique', Journées d'études sur les systèmes d'information élaborée de la SFBA, Ile Rousse, Corse, 1997

[DUCLOY 91] DUCLOY J., CHARPENTIER P., FRANCOIS C., GRIVEL L. (1991) "Une boîte à outils pour le traitement de l'Information Scientifique et Technique", 4es. Journées Internationales Le Génie logiciel et ses applications. Toulouse, 9-13 Décembre 1991, p. 239-254 ; et dans Génie logiciel, n° 25, 1991, p. 80-90.

[DUCLOY 99] DUCLOY J.,'DILIB, une plate-forme XML pour la génération de serveurs WWW et la veille scientifique et technique, Le Micro Bulletin Thématique n°3, L'information scientifique et technique et l'outil Internet, Editeur CNRS-DSI, 1999, p.113-137.

[DUCOURNEAU 98] Langages et modèles et objets, Editeurs DUCOURNEAU R. EUZENAT J. MASINI G. NAPOLI A . Collection Didactique, INRIA, 527 p.

[DUSOULIER 91] DUSOULIER N., DUCLOY J. "Processing of data and exchange of records in a scientific and technical information center. Formats : what for ?" UNIMARC/CCF Workshop, Florence (IT) (IFLA/UNESCO), 05-07 Juin 1991

[FERNANDEZ 93] FERNANDEZ M.T., CABRERO A., ZULUETA M.A., GOMEZ T. 'Constructing a relational database for bibliometric analysis', Research Evaluation, 1993, Vol 3,n°1, 55-62.

[FAUCOMPRE 98] FAUCOMPRE P. 'La mise en correspondance automatique de banques de données bibliographiques scientifiques et techniques à l'aide de la classification internationale de brevets'. Thèse de doctorat en Sciences de l'information et de la communication. Université Aix Marseille III, 1998.

[GLANZEL 96] GLÄNZEL W. 'The Need for Standards in Bibliometric Research and Technology', Scientometrics, vol.35, N°2 (1996) , 167-176.

[GOLDFARB 90] GOLDFARB C. *The SGML Handbook*, Oxford, Oxford University Press. (1990)

[GOMEZ 96] GOMEZ I., BORDONS M., FERNANDEZ M.T., MENDEZ A. 'Copying with the problem of Subject Classification Diversity', Scientometrics, , vol.35, N°2 (1996), 223-236.

[GRIVEL 95] GRIVEL L., FRANÇOIS C. Conception et développement d'un système d'information dédié à la veille scientifique, basé sur les sorties des outils de classification thématique : SDOC et NEURODOC , In : BALPE J.P, LELU A., SALEH I,.Eds, *Hypertexte et hypermedia, réalisations, outils et méthodes*, Paris, Editions Hermès: 109-118.

[GRIVEL 95b] GRIVEL L., FRANÇOIS C. "Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et

technique", *SOLARIS* n° 2, Presses Universitaires de Rennes, p. 81-112 (1995); et dans http//www.info.unicaen/bnum/jelec/Solaris.
[GRIVEL 97] GRIVEL L., POLANCO X., KAPLAN A. 'A computer system for big scientometrics at the age of the World Wide Web', Scientometrics, vol.40, N°3 (1997), 493-506
[GRIVEL 99] GRIVEL L. 'HENOCH, un outil d'analyse de corpus d'information scientifique et technique', Le Micro Bulletin Thématique n°3, L'information scientifique et technique et l'outil Internet, Editeur CNRS-DSI, 1999, p.27-44.
[HERWIJNEN 90] HERWIJNEN E. "Practical SGML", Kluwer Academic Publishers, 1990
ISO 8879 - 1986. Information processing - Text and office systems - Standard Generalised Markup Language (SGML), 155 pages
[MICHARD 98] MICHARD A. 'XML Langage et application' Editions Eyrolles, 361 p, 1998
[MOED 88] MOED H.F 'The use of On-line databases for bibliometric analysis', In L. Egghe and R. Rousseau (editors), Informetrics 87/88 (Elsevier Science Publishers), Amsterdam), 145-158
[MOED 95] MOED H.F, DE BRUIN R.E, Van LEEUWEN TH. 'New bibliometric tools for the assessment of National Research Performance : Database description, overview of indicators and first applications', Scientometrics, Vol.33, n°3 (1995), 381-422.
[MOED 95b] MOED H.F, Van LEEUWEN TH. 'Improving th accuracy of the ISI's journal impact factor, Journal of the American Society for Information Science, 46 (1995), 381-422.
[MOED 96] MOED H.F. 'Differences in the construction of SCI Based Bibliometric Indicators among Various Producer : A first Overview' , Scientometrics, , vol.35, N°2 (1996), 177-192
[NAUER 99] NAUER E. 'De l'importance de la normalisation en bibliométrie', Journées d'études sur les systèmes d'information élaborée de la SFBA, Ile Rousse, Corse, 27 septembre-1er octobre 1999
[POLANCO 95] POLANCO X. 'Aux sources de la scientométrie', in : SOLARIS, «Les sciences de l'information : bibliométrie, scientométrie, infométrie, sous la direction de Jean-Max Noyer ». Edition : Presses Universitaires de Rennes, 1995, pp.13-78.
[ROYAUTE 99] ROYAUTE J. Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information, Thèse de doctorat en informatique, Université H. Poincaré Nancy I, 1999.
[RAE 97] Rapport européen sur les indicateurs scientifiques et technologiques 1997, Annexes méthodologiques, note méthodologique D.
[Rapport OST 1998] Science et Technologie Indicateurs 1998, annexes méthodologiques
[ROSTAING 96] ROSTAING H. 'La bibliométrie et ses techniques', Edition : sciences de la sociéte, coll : « Outils et méthodes », 1996, 131p.
[SMALL 95] SMALL H. 'Relational bibliometrics', In: Michael E.D Koenig, Abraham Bookstein (Eds), 5th International Conference on Scientometrics and Informetrics, Learned Information Inc. Medford NJ, 525-530.
[VINKLER 96] VINKLER P. 'Standardization of Scientometric Indicators', vol.35, N°2 (1996), 237-245.
[ZITT 96] ZITT M. , TEIXEIRA N. 'Science Macro-Indicators : some aspects of OST Experience Scientometrics', vol.35, N°2 (1996), 209-222.

Annexe 1 : une notice extraite du SCIENCE CITATION INDEX (SERVEUR : Dialog)

Nb de réf citées

Laboratoires

20/5/1
12508319 Genuine Article#: LT747 Number of References: 52
Title: AMYOTROPHIC-LATERAL-SCLEROSIS AND STRUCTURAL DEFECTS IN CU,ZN SUPEROXIDE-DISMUTASE
Author(s): DENG HX; HENTATI A; TAINER JA; IQBAL Z; CAYABYAB A; HUNG WY; GETZOFF ED; HU P; HERZFELDT B; ROOS RP; WARNER C; DENG G; SORIANO E; SMYTH C; PARGE HE; AHMED A; ROSES AD; HALLEWELL RA; PERICAKVANCE MA; SIDDIQUE T
Corporate Source: NORTHWESTERN UNIV,SCH MED,DEPT NEUROL,300 E SUPER ST/CHICAGO//IL/60611; NORTHWESTERN UNIV,SCH MED,DEPT NEUROL,300 E SUPER ST/CHICAGO//IL/60611; SCRIPPS CLIN & RES INST,DEPT MOLEC BIOL/LA JOLLA//CA/92037; NORTHWESTERN UNIV,INST NEUROSCI/CHICAGO//IL/60611; UNIV CHICAGO,DEPT NEUROL/CHICAGO//IL/60637; DENT NEUROL INST,DEPT NEUROL/BUFFALO//NY/14209; DUKE UNIV,MED CTR,DEPT MED NEUROL/DURHAM//NC/27710; UNIV LONDON IMPERIAL COLL SCI TECHNOL & MED,DEPT BIOCHEM/LONDON SW7 2AZ//ENGLAND/; NORTHWESTERN UNIV,SCH MED,DEPT CELL MOLEC & STRUCT BIOL/CHICAGO//IL/60611

Source

Catégorie de **périodique** (et non plan de classement)

Journal: SCIENCE, 1993, V261, N5124 (AUG 20), P1047-1051
ISSN: 0036-8075
Language: ENGLISH Document Type: ARTICLE
Geographic Location: ENGLAND; USA
Subfile: SciSearch; CC PHYS--Current Contents, Physical, Chemical & Earth Sciences; CC LIFE--Current Contents, Life Sciences; CC AGRI--Current Contents, Agriculture, Biology & Environmental Sciences
Journal Subject Category: MULTIDISCIPLINARY SCIENCES
Abstract: Single-site mutants in the Cu,Zn superoxide dismutase (SOD) gene (SOD1) occur in patients with the fatal neurodegenerative disorder familial amyotrophic lateral sclerosis (FALS). Complete screening of the SOD1 coding region revealed that the mutation Ala4 to Val in exon 1 was the most frequent one; mutations were identified in exons 2, 4, and 5 but not in the active site region formed by exon 3. The 2.4 angstrom crystal structure of human SOD, along with two other SOD structures, established that all 12 observed FALS mutant sites alter conserved interactions critical to the beta-barrel fold
and dimer contact, rather than catalysis. Red cells from heterozygotes had less than 50 percent normal SOD activity, consistent with a structurally defective SOD dimer. Thus, defective SOD is linked to motor neuron death and carries implications for understanding and possible treatment of FALS.

Mots-clés obtenus par indexation automatique

Identifiers--KeyWords Plus: MANGANESE; PROTEIN; ENZYME; MUTATIONS; INTERFACE; STABILITY; DISEASE; LINKAGE
Research Fronts: 91-2104 002 (SUPEROXIDE DISMUTASES; REACTIVE OXYGEN SPECIES; ANTIOXIDANT ENZYMES)
91-0391 001 (ENDOTHELIUM-DERIVED RELAXING FACTOR NITRIC-OXIDE SYNTHASE; L-ARGININE PATHWAY; CONTINUOUS BASAL EDRF RELEASE)
91-1725 001 (CU,ZN SUPEROXIDE-DISMUTASE ACTIVITY; COPPER SITES; INACTIVE PROENZYME IN ANAEROBIC YEAST)
91-2496 001 (2.5-A RESOLUTION; CRYSTAL-STRUCTURE OF MANDELATE RACEMASE; TRYPANOSOMAL TRIOSEPHOSPHATE ISOMERASE; CRYSTALLOGRAPHIC REFINEMENT)
91-3964 001 (POLYMERASE CHAIN-REACTION; FACTOR-IX GENE; SEVERE HEMOPHILIA-B HAVING A POINT MUTATION; RAPID DETECTION OF SINGLE BASE MISMATCHES; DYSTROPHIN MESSENGER-RNA)
91-4514 001 (2.4-A RESOLUTION; MOLECULAR REPLACEMENT; X-RAY CRYSTALLOGRAPHY ANALYSIS; BOVINE PANCREATIC TRYPSIN-INHIBITOR; NERVE GROWTH-FACTOR)
91-4817 001 (LIPASE GENE; CDNA FOR STIMULATORY GDP/GTP EXCHANGE PROTEIN; EXPRESSION OF MESSENGER-RNA)
91-6189 001 (BRAIN SUPEROXIDE-DISMUTASE ACTIVITY FOLLOWING FOREBRAIN ISCHEMIA IN RAT; REACTIVE OXYGEN SPECIES; NERVE GROWTH-FACTOR; INVIVO GENERATION)
Cited References:
ANTONARAKIS SE, 1992, V14, P1126, GENOMICS
BEAUCHAMP CO, 1971, V44, P276, ANAL BIOCHEM

Références citées (format : ordre alphabétique, 1er auteur, année, volume, 1ère page, titre périodique abrégé)

.../...

Annexe 2 : Tableau comparatif des trois organismes étudiés

	OST Observatoire des Sciences et des Techniques 93, rue de Vaugirard 75006 PARIS Tél. : 01 42 22 30 30 Télécopie : 01 45 48 63 94	**CINDOC** Centro de Informacion y Documentacion Cientifica Joaquin Costa 22 28002 Madrid Tél : +34-1-5635482 Télécopie : +34-1-5642644	Centre f Leiden U PO Box 2300 RE Tel : +3 Fax : +3
Missions	« construire des indicateurs fiables, pertinents et pérennes, décrivant la science et la technologie françaises en comparaison européenne et internationale »	« élaboration de bases de données bibliographiques et réalisation d'analyses bibliométriques de la production scientifique espagnole, ainsi que normalisation de la terminologie scientifique »	cartogra particuli méthode bibliomé
Type d'organisme et effectif	Groupement d'Intérêt Public (GIP) de 14 membres : 7 ministères, 6 grands établissements publics (CEA, CNRS, CNES, CNET, INSERM, INRA) et l'ANRT. Membre associé : ORSTOM effectif environ 10 personnes	Centre de documentation scientifique du CSIC, (Consejo Superior de Investigaciones Cientifica). Environ 130 personnes	Centre Organiz 8 cherch
Produits de l'organisme	***Publications*** : Indicateurs science et technologie, rapports annuels. La lettre de l'OST Les cahiers de l'OST ***Produits des ateliers de l'OST*** pour analyse stratégique à la demande (micro-indicateurs).	services comparables à ceux de l'INIST (fourniture de documents, recherches bibliographiques, traductions...), bases de données multidisciplinaires ICYT (science et technique) et ISOC (sciences humaines). Toutes ces bases de données couvrent spécifiquement la littérature espagnole. ***Concernant l'Infométrie*** - une base de données bibliométrique - une revue électronique *Cybermetrics* : journal international de recherche en scientométrie, bibliométrie et infométrie.	- une ba ***Publica*** - article - rappor comman culture e des Pays - partici deux ans

Ressources	Pour calculer les indicateurs *bibliométriques* standards en sciences et techniques • les données du Science Citation Index (SCI), après extraction de certains journaux de psychologie et d'économie, enrichissement avec Compumath, produite elle aussi par l'ISI. • les bases EPAT et USPAT (brevets européens et américains enquêtes ministérielles, R.D. (recherche industrielle et innovation), MENDEP (étudiants et diplômés), OCDE, UNESCO, EUROSTAT (statistiques européennes), bases de données bibliographiques (PASCAL INSPEC, CHEMICAL ABSTRACT, SCI)	• Des bases de données bibliographiques (SCI, SSCI, ICYT, Physic Brief, INSPEC, Chemical Abstract, Biosis, MEDLINE, Exerpta Medica). • Des données factuelles : rapports officiels annuels et données de ressources humaines du monde scientifique et universitaire espagnol	Une bas constitue chercheu pour SC Science Humanit (Institute S'ajoute données chercheu
Types d'indicateurs	*MACROINDICATEURS* : niveau d'observation à un niveau agrégé (pays, région), en comparaison internationale • mesure de niveau d'activité • indicateurs de spécialisation • indicateurs d'impacts • profils d'activité • copublications • cocitations • codépôt de brevet • matrices inventeurs-déposants de brevets *MICROINDICATEURS* : ciblés sur le plan géographique, institutionnel,	Macroindicateurs d'impact : Espagne en comparaison internationale • IF : Facteur d'impact moyen (pour une spécialité au niveau national) • RIF : Relative Impact Factor (comparaison internationale) Microindicateurs d'impact : comparaison des différents centres de recherches dans la même discipline Indicateurs de production scientifique par spécialité. Indicateurs de production scientifique par lieu. Copublications par spécialité. Copublications par lieu.	Sept typ 1) Des in 2) Des in 3) Des in différent 4) Des in 5) Des in collabor 6) Des in 7) Des in (revues s

	produits à la demande		

Annexe 3
Le couplage SGML/SGBD pour la fusion de données multi-sources

1 Description d' HENOCH

Le système HENOCH comprend:
1. un générateur de bases de données relationnelles à partir de documents au format SGML. Ce générateur utilise la notion d'arbre SGML comme structure pivot pour la description des données alimentant la base infométrique. Ces documents sont :

a) les données initiales (qui sont de différents types et qui peuvent provenir de différentes sources : articles de périodiques, congrès, thèses, brevets) mises au format SGML et complétées (éventuellement) d'un certain nombre d'informations obtenues par traitements linguistiques (mot clés)

b) les résultats de classification des données initiales (regroupement de documents ou de mots-clés) par les outils SDOC et NEURODOC [GRIVEL 95b],

c) les tables de nomenclatures nécessaires pour la production de certains indicateurs.

2. un générateur des systèmes hypertextes sous WWW pour l'analyse, la valorisation et la diffusion des résultats de classification. Ce programme établit une interface WWW-SGBD par une passerelle qui permet de se connecter au SGBD, soumettre des requêtes SQL à partir d'un modèle de page HTML incluant des requêtes SQL, récupérer le résultat et le mettre au format HTML conformément au modèle, et enfin se déconnecter.

Le générateur de base relationnelle procède en deux étapes :

1) Création du 'squelette' de la base selon un modèle de données suffisamment générique pour prendre en compte la diversité des types de documents

Le 'squelette' de la base correspond à la définition de l'ensemble des tables utilisées (nom de la table, attributs, type de chaque attribut).

2) Analyse des documents SGML et chargement des données dans la base

Pour chaque type de document au format SGML, un fichier de configuration basé sur un modèle de description de document (Document Type Definition DTD) permet d'associer un traitement (par exemple, tous les tests à effectuer avant d'insérer des valeurs dans la table) à un ou plusieurs éléments de données pour chaque table pour assurer la cohérence des données dans la base. Ces procédures, écrites en PL-SQL, sont stockées dans la base.
L'appel aux procédures d'insertion s'effectue donc lors de l'analyse du document SGML par un parser (analyseur syntaxique) qui, à partir d'un fichier de configuration, associe le contenu de chaque balise avec chaque attribut de chaque table.

2 La fusion de données multi-sources

L'idée est de prendre le meilleur de chacune des sources dans son format initial. Au lieu d'éliminer les doublons en ne gardant qu'un exemplaire de notice pour chaque clé, en privilégiant par exemple un ordre de préférence dépendant de la base d'origine [NAUER 99], les doublons sont ici considérés comme sources de richesses pour construire des « super-notices », via des requêtes SQL, en prenant par exemple, tel champ d'une source et tel autre

d'une autre source, ou en combinant deux champs, sur la base de la présence ou de l'absence de telle ou telle information.

Il est en effet possible de mettre en place une procédure de repérage du même article dans les différentes sources (dédoublonnage) puis de s'appuyer sur le modèle relationnel pour combiner les informations provenant des différentes sources en vue de constituer des descriptions d'unités documentaires les plus complètes possibles en retenant le 'meilleur' des différentes bases.

Pour cela, chaque document est identifié par une clé unique construite à partir de différents éléments de données (auteurs, année de publication, etc.). Avant de créer un nouvel enregistrement dans la table des documents, la procédure d'insertion récupère chacun des éléments de données nécessaire à la construction de la clé et vérifie l'absence de cette clé dans la table. Si c'est le cas, un numéro unique (NuméroDocument) est attribué au document. Les documents ayant la même clé ont le même numéro de document.

Puis chaque élément d'information (titre de périodique, auteur, etc.) du document alimente la table lui correspondant (table des périodiques, table des auteurs, etc.) en lui associant le numéro de document correspondant.

La « reconstitution » du document sous forme de super-notice est effectuée par jointure sur le numéro identifiant le document entre toutes les tables (auteur, pays, titre de journal, etc.).

Le résultat de cette requête peut alors être exporté par le générateur d'hypertexte sous forme de données XML pour des traitements ultérieurs ou pour être accessible par un browser.

L'intérêt de cette architecture est la simplicité avec laquelle il est possible de fusionner des données provenant de plusieurs base hétérogènes et de définir un formatage global cohérent pour le résultat formé par l'ensemble des données fusionnées.