Construction d'une base de données bibliométrique, résultant de la fusion des bases bibliographiques Pascal et Science Citation Index, en vue de l'élaboration d'indicateurs thématiques pour le CNRS

DASSA Michèle (*) et POUPON-CZYSZ Catherine (**)
michele.dassa@cnrs-dir.fr,czysz@inist.fr

(*) CNRS - UNIPS - 3, rue Michel-Ange 75794 PARIS CEDEX 16. (France)
(**) INIST - CNRS - 2, allée du Parc de Brabois 54514 VANDOEUVRE-LES-NANCY CEDEX. (France)

Mots-clés :
CNRS - INIST - Bibliométrie – Méthodologie - Base de données Pascal - Base de données SCI – Publication scientifique – Langage documentaire

Keywords :
CNRS - INIST – Bibliometrics– Methodology – Pascal Database - SCI Database – Scientific publication – Information language

Palabras clave :
CNRS - INIST – Bibliometrìa– Metodologìa – Base-dato Pascal – Base-dato SCI– Publicacciòn cientìfica – Lenguaje-documental

Résumé
En 2003, dans le cadre du contrat d'action pluriannuel CNRS-Etat, l'Unité d'Indicateurs de politique Scientifique (UNIPS) du CNRS a engagé un travail en partenariat avec l'Institut de l'Information Scientifique et Technique (INIST) pour fournir à la direction du CNRS des indicateurs bibliométriques permettant de suivre l'évolution des cinq grands axes interdisciplinaires prioritaires.
L'UNIPS a construit une base bibliométrique à usage interne, à partir de la fusion des données issues de la base Science Citation Index (SCI), produite par l'Institute for Scientific Information (ISI) et de la base Pascal de l'INIST. Elle permettra de bénéficier des qualités de la première pour la bibliométrie (comptage des citations) et de la richesse de l'information scientifique pour la seconde (descripteurs, codes de classement).
Cette étude s'est déroulée en plusieurs étapes, dont les deux premières font l'objet de cette communication car détaillant une démarche peu étudiée :

- ✓ formulation en stratégies de recherche des cinq axes prioritaires pour extraction des références bibliographiques de Pascal
- ✓ fusion des données issues des bases Pascal et SCI, afin d'en exploiter les données à valeur ajoutée ;
- ✓ production des indicateurs bibliométriques.

Cette communication présente la méthodologie utilisée, les problèmes rencontrés et les résultats obtenus.

1. Introduction

En 2003, dans le cadre du contrat d'action pluriannuel (CAP) CNRS-Etat 2002-2005 [1], l'Unité d'Indicateurs de Politique Scientifique (UNIPS) [2] a engagé un travail en partenariat avec le service Veille de l'Institut de l'Information Scientifique et Technique (INIST) [3] pour fournir à la direction du CNRS des indicateurs bibliométriques permettant de suivre l'évolution des cinq grands axes interdisciplinaires prioritaires de 1996 à 2001 ("Le vivant et ses enjeux sociaux", "Information, communication et connaissance", "Environnement, énergie et développement durable", "Nanosciences, nanotechnologies et nanomatériaux", "Astroparticules : des particules à l'Univers").

Pour produire les indicateurs, les bases de données bibliographiques suivantes ont été utilisées :

- ✓ Pascal, base de données produite par l'INIST, offrant une grande qualité d'indexation ;
- ✓ Science Citation Index (SCI), base produite par l'Institute for Scientific Information (ISI), référence internationale pour les études bibliométriques, mais sans classification thématique fine.

SCI est devenue en quelques années la base de données de référence pour les études bibliométriques. Elle répertorie des périodiques scientifiques "cœur", c'est-à-dire des revues à comité de lecture international et ayant le plus fort facteur d'impact [4]. Ces périodiques sont considérés comme représentatifs de la production mondiale de bon niveau [1].
La base SCI présente cependant un certain nombre de limites explicitées par différents auteurs ([1], [2]) :

- ✓ sur-représentation des revues américaines et plus généralement anglo-saxonnes,
- ✓ mauvaise couverture de certains domaines (comme les Mathématiques),
- ✓ absence d'indexation fine au niveau des articles.

Pour produire les indicateurs bibliométriques dans le cadre du contrat d'action pluriannuel, il est nécessaire de caractériser les publications des unités CNRS en fonction de thématiques précises. Or la base SCI ne dispose pas d'indexation précise. Il est seulement possible d'attribuer, à un article, un domaine défini par le périodique dans lequel il est publié. Mais cette classification au niveau d'un périodique ne peut être aussi pertinente qu'une classification effectuée au niveau de l'article. L'Observatoire des sciences et techniques (OST) [5] a construit une classification en huit domaines, fondée sur la classification de périodiques de l'ISI et utilisée par l'UNIPS pour ses études bibliométriques [8].

Ainsi, les limites de la base SCI et les besoins et intérêts spécifiques du CNRS nous ont-ils amenés à réfléchir à la mise en place d'une nouvelle base de données "multibase" à usage bibliométrique et permettant des comparaisons au niveau international.
Certains observatoires (en France, en Espagne, en Hollande, au Canada) produisant des indicateurs bibliométriques ont abouti à la même constatation ([3], [4], [6]).
Ces observatoires ont utilisé la base SCI complétée avec d'autres bases ou sources de données.
Par exemple :

- ✓ l'OST, utilise la base SCI modifiée et complétée par la base Compumath [1] ;
- ✓ l'Observatoire des Sciences et Technologies au Canada (CIRST) utilise une base construite à partir du SCI, du Social Science Citation Index et du Art & Humanities Citation Index [3].

---

[1] Pour plus de détails, voir : http://www.cnrs.fr/cw/fr/accu/contratPluri/CAP.htm
[2] UPS78 : Unité propre de service du CNRS
[3] UPS76 : Unité propre de service du CNRS
[4] Le facteur d'impact est le rapport, sur une période de 2 ans, du nombre de citations sur le nombre d'articles publiés. Il est donné dans le Journal of Citation Report (JCR), base de données produite par l'ISI, analysant environ 5 680 périodiques scientifiques et techniques couvrant toutes les disciplines.
[5] Pour plus de détails, voir : http://www.obs-ost.fr/fr/

L'utilisation de Pascal en complément de la base SCI pour réaliser des indicateurs bibliométriques a déjà été étudiée ([3], [5], [7]). Cependant, son utilisation conjointement au SCI est restée marginale et n'a jamais été utilisée en production.

Le choix de Pascal s'est imposé : base multidisciplinaire produite par le CNRS et présentant un volume du même ordre que SCI avec une indexation au niveau de chaque article.

L'exploitation de données hétérogènes est un enjeu de plus en plus important.
Pour mettre en relation différentes sources de données bibliographiques, de nombreuses approches sont possibles : réalisation de liens sémantiques entre les bases [9], réalisation d'interfaces d'interrogation ...
Nous avons choisi de réaliser la liaison entre Pascal et SCI en effectuant une jointure article par article. Ceci implique, en même temps, la formulation des domaines sémantiques à étudier en requêtes à effectuer sur Pascal. La production d'indicateurs bibliométriques, par cette approche, nécessite 3 phases indissociables :

- ✓ formulation en stratégies de recherche des cinq axes prioritaires pour extraction des références bibliographiques de Pascal ;
- ✓ fusion des données issues des bases Pascal et SCI, afin d'utiliser simultanément les informations à valeur ajoutée contenues dans les deux bases ;
- ✓ production des indicateurs bibliométriques (comme par exemple le nombre de publications scientifiques des unités CNRS rapporté à la France, à l'Europe et au Monde).

Les deux premières étapes font l'objet de cette communication.
La production des indicateurs bibliométriques de suivi et d'objectifs du contrat d'action pluriannuel (troisième étape) a été réalisée au cours de l'année 2004, pour les années de publication 1996 à 2001. Ces indicateurs ont été présentés au conseil d'administration du CNRS le 24 juin 2004. Ils sont disponibles sur l'intranet du CNRS (accessible par tous les laboratoires CNRS).
L'objectif principal de cette étude est de pouvoir relier des thématiques précises (de la base Pascal) à des données bibliométriques (issues de la base SCI).
Dans cet article, nous souhaitons mettre en évidence les points clés et les difficultés rencontrées pour construire, utiliser cette nouvelle base de données bibliométrique et évaluer sa qualité.

2. Matériel et méthodes

2.1. Choix des bases sources

2.1.1. Pascal [6]

- ✓ Avantages : cette base est reconnue pour son caractère multidisciplinaire, sa couverture de la littérature mondiale (3 868 titres vivants analysés dans Pascal), avec une place importante accordée à la littérature française et européenne (45% des documents signalés dans Pascal). Elle offre des possibilités d'interrogation par thématique (à partir des codes d'un plan de classement hiérarchique [7]) ou par descripteurs. 88% des documents analysés sont des périodiques, 3% des thèses, monographies et rapports, et 9% des congrès. A titre d'exemple, pour l'année de publication 1999, environ 500 000 références bibliographiques, dont 465 000 issues de périodiques ont été analysées.

- ✓ Inconvénients : le catalogage automatisé des affiliations, sans traitement d'homogénéisation des noms d'organisme, ne permet pas un repérage simple des laboratoires CNRS.

[6] Base de données bibliographiques en Science, Technologie et Médecine.
Pour plus de détails, voir : http://www.inist.fr/PRODUITS/pascal.php
[7] Il permet d'attribuer un code alphanumérique aux notices et de classer ainsi les articles en fonction du sujet.
Pour plus de détails, voir : http://connectsciences.inist.fr/bases/internes/plan_classement/resdoc_planclass.php

En effet, plusieurs cas de figure se présentent :

- la mention claire de l'appartenance au CNRS est présente mais prend différentes formes : CNRS, C.N.R.S, Cent. Nat. de Rech. Scient.... ;
- la mention d'appartenance au CNRS est indirecte, faite par le code de la structure : UMR, UPR, UPS, UMS, FRC, FR, IFR, FRE, GDR, MOY... ;
- la mention d'appartenance au CNRS est absente, seul le nom du laboratoire (libellés longs, ou formes abrégés variables) et/ou son sigle sont présents.

2.1.2. Science Citation Index [8]

✓ Avantages : cette base répertorie chaque année la plupart des publications mondiales avec une forte prédominance de littérature anglo-saxonne. Elle présente des qualités pour les études bibliométriques (exactitude de l'enregistrement des adresses, comptage des citations...). A titre d'exemple, pour l'année de publication 1999, cette base contient environ 800 000 références bibliographiques issues de quelques 3 600 périodiques.

✓ Inconvénients : elle ne possède pas de classification thématique fine et aucune monographie n'est analysée.

Afin de construire des indicateurs bibliométriques sur les publications scientifiques des unités du CNRS, l'UNIPS complète les données d'affiliation de SCI (champ "corporate source") par analyse semi-manuelle des enregistrements d'adresses (plus de 75 000 pour l'année 2000).
Une publication scientifique d'unité CNRS est une publication dont l'un des auteurs, qu'il soit salarié du CNRS ou non, a donné comme adresse une unité soutenue et évaluée par le CNRS (laboratoire propre, mixte ou associé). A contrario, les publications signées par des chercheurs du CNRS qui travaillent dans des unités strictement INSERM ou universitaires (non mixtes avec le CNRS) ne sont pas reconnues comme provenant des laboratoires CNRS.

2.2. Spécification du format des données et formulation des axes du CAP

2.2.1. Spécification du format des données

La base de données Pascal est stockée à l'INIST en format SGML [9].
Afin de pouvoir exploiter ces données avec ses outils, l'UNIPS a demandé un reformatage de SGML vers un format simplifié de type serveur (cf ci-dessous).

Exemple d'une référence en SGML (extrait) avant et après reformatage en format de type serveur :

Format SGML

```
…<fC01 dir="010FRE">
  <s0>Si le bois est un des plus anciens mat&eacute;riaux... l'attaque des insectes dans leurs r&eacute;gions.</s0>
</fC01>
<fC02 dir="01X000">[10]
        <s0>001D14H02B</s0>
</fC02>
<fC02 dir="02X000">
        <s0>001D14E02</s0>
</fC02>
```

[8] Pour plus de détails, voir : http://www.isinet.com
[9] Standard Generalized Markup Language. Métalangage utilisé pour définir de façon générale des langages permettant de structurer des documents sous forme d'arbre (ISO 8879). XML en est un dérivé.
[10] La zone répétitive <fC02> correspond aux codes de classement.

<fC02 dir="03X000">
<s0>295</s0>
</fC02>
<fC03 dir="01XFRE">[11]
<s0>Construction bois</s0>
<s5>01</s5>
</fC03>
…

Format simplifié de type serveur
AN: 01-0342620
TI: La lutte contre les termites
TT: The fight against the termites
DT: Publication en série; Niveau analytique
JN: Batiment information : (Clichy)
IS: 1266-8176
NO: 43
PG: p.11
LA: Français
AF: Si le bois est un des plus anciens matériaux de construction... des insectes dans leurs régions.
CC[12]: 001D14H02B; 001D14E02; 295
DS[13]: Construction bois; Altération matériau; Biodétérioration; Termite;
Répartition géographique; Infestation; Réglementation; Bâtiment; Insecta; Prévention; Site Web; Internet; France
TG[14]: Arthropoda; Invertebrata; Europe
PY: 2001

Au final, l'intégralité de la base Pascal, de 1996 à 2003, a été traitée et fournie à l'UNIPS, soit un total d'environ 3 millions de références bibliographiques.

2.2.2. Formulation des axes prioritaires du CAP

La formulation en requêtes des cinq axes cités précédemment a été une opération complexe et longue, réalisée par des ingénieurs documentalistes de l'INIST, spécialistes scientifiques des domaines abordés. Elle a nécessité, pour certains axes, des précisions supplémentaires (termes pas assez explicites,…) par rapport au document de base[15].

Il est à signaler que les libellés des domaines et sous-domaines de Pascal (donnés par le plan de classement hiérarchique) sont sans rapport avec les axes scientifiques décrits dans le CAP.

<table>
<tr><th>Domaines</th><th colspan="2">Sciences exactes et Technologie (001)</th><th colspan="2">Sciences biologiques et médicales (002)</th></tr>
<tr><td rowspan="5">Sous-Domaines</td><td>Sciences et Techniques communes</td><td>001A</td><td rowspan="2">Sciences biologiques fondamentales et appliquées. Psychologie</td><td rowspan="2">002A</td></tr>
<tr><td>Physique</td><td>001B</td></tr>
<tr><td>Chimie</td><td>001C</td><td rowspan="3">Sciences médicales</td><td rowspan="3">002B</td></tr>
<tr><td>Sciences appliquées</td><td>001D</td></tr>
<tr><td>Terre, Océan, Espace</td><td>001E</td></tr>
</table>

Une même référence peut se voir attribuer plusieurs codes de classement, marquant ainsi le caractère interdisciplinaire des articles.

[11] La zone répétitive <fC03> correspond aux descripteurs spécifiques.
[12] Le champ CC correspond aux codes de classement.
[13] Le champ DS correspond aux descripteurs spécifiques.
[14] Le champs TG correspond aux termes génériques.
[15] http://www.cnrs.fr/cw/fr/accu/contratPluri/CAP.htm

Les stratégies de recherche ont parfois nécessité la combinaison complexe (avec plusieurs types d'opérateurs booléens) de codes de classement, de descripteurs français et anglais, de termes contenus dans le titre, de nom de périodiques. Pour l'axe " Nanosciences, nanotechnologies et nanomatériaux", environ 50 mots clés ont été combinés.

L'INIST a fourni, pour les cinq axes de recherche, une liste de plusieurs dizaines de milliers de clés permettant d'identifier les références concernées parmi les données Pascal données à l'UNIPS.

3. Fusion des bases de données et appariement des données

3.1. Intégration des données Pascal dans une base de données SQL-Server

Afin de pouvoir réaliser les traitements, des bases de données relationnelles sur un serveur de l'UNIPS ont été créées en éclatant les informations intéressantes dans différents champs. Nous avons choisi de construire une table par paramètre (année de publication, titre, ISSN ...). Chaque référence est identifiée par un numéro unique, ce qui permet ensuite de réaliser des liens sans ambiguïté. Par exemple, cela permet de retrouver, pour chaque article, son titre, l'année de publication ou l'ISSN du périodique dans lequel il est publié.
Le logiciel "Access" est utilisé comme requêteur.

3.2. Appariement automatique des données Pascal avec les données SCI

3.2.1. Précision concernant l'année de publication

Pour les bases de données utilisées, une forte proportion d'articles (près de 10% pour SCI et 35% pour Pascal) est cataloguée l'année qui suit leur publication. Par conséquent, pour chaque année de publication, les calculs seront réalisés sur un corpus de références bibliographiques traitées l'année de leur publication et l'année suivante. En effet, compte-tenu des volumes de données traités, les corpus de références Pascal ont été constitués par année de chargement dans la base, et non pas par année de publication.

3.2.2. Choix des critères d'appariement

Une étude préalable a été nécessaire pour définir le meilleur compromis dans le choix des champs à apparier. Ce choix est très important car il détermine la qualité des résultats, en effet :

- ✓ Si le nombre de critères est insuffisant, des références bibliographiques peuvent être associées à tort (trop de bruit). C'est le cas si l'on choisit de joindre des références d'articles publiés la même année, dans le même périodique mais qui auraient une pagination différente (année de publication et ISSN identiques uniquement).
- ✓ Si les critères sont trop nombreux, des références ne seront pas associées alors qu'elles sont identiques (trop de silence). C'est le cas par exemple, si l'on souhaite joindre des références d'articles qui ont exactement les mêmes auteurs. Les noms et les prénoms des auteurs présentent des formes trop hétérogènes d'une base à une autre.

Le compromis est aussi celui choisi par l'OST lors d'une étude de 1998 sur la comparaison des bases SCI et Pascal [7].

Les critères retenus sont :

- ✓ l'ISSN[16] du périodique ;
- ✓ l'année de publication du périodique ;
- ✓ le volume du périodique ;
- ✓ la première page de l'article.

3.2.3. Appariement automatique des données Pascal et SCI

Une référence contenue dans Pascal sera donc considérée comme identique à une référence de SCI si :

année de publication, ISSN, volume, page de début (SCI) = année de publication, ISSN, volume, page de début (Pascal)

L'année de publication est codée sur 4 chiffres, l'ISSN sur 8 caractères, la première page sur 4 chiffres (p.20 sera traduit en 0020) et le volume sur 4 chiffres (0315).

Une étude préalable nécessaire a identifié tous les cas de figures pour lesquels la transformation était automatisable. Sur un échantillon de notices, a été vérifiée l'identité, dans les deux bases, du titre et de la pagination de l'article, du volume et de la date de publication du périodique.
La transformation automatique a dû être complétée par un travail manuel pour certaines configurations.

3.3. Difficultés rencontrées

3.3.1. L'ISSN du périodique

Les notices de la base SCI ne possèdent pas d'ISSN. Il a été préalablement nécessaire d'associer un ISSN à un périodique. Nous avons récupéré, sur le site de l'ISI[17], la liste des périodiques (et leur ISSN) analysés dans la base SCI. Cette liste a été complétée car un certain nombre de périodiques, répertoriés par exemple en 1999, ne sont plus analysés en 2003 ou bien ont changé d'ISSN. Nous avons pu retrouver les ISSN de tous les périodiques.
Dans un deuxième temps, un contrôle supplémentaire automatique puis manuel a été effectué pour repérer les périodiques qui avaient dans les deux bases le même titre mais des ISSN différents. En effet, l'INIST met à jour ou utilise l'ISSN validé par le centre international de l'ISSN[18], alors que l'ISI ne catalogue que l'ISSN apparaissant sur le périodique (même si non validé). Pour chaque année, après comparaison des données, les articles d'une quinzaine de périodiques ont ainsi pu fusionner.
Pour l'année de publication 2001, nous nous sommes heurtés à un nouveau problème : celui des périodiques qui existent à la fois sous forme papier et électronique, et qui conservent la même pagination mais ont des ISSN différents. En 2001, ce problème est marginal (concerne moins d'une dizaine de revues) mais il pourrait s'accentuer dans les prochaines années.

3.3.2. L'année de publication du périodique

Les données de ce champ ont dû être reformatées car c'est la date de publication qui est indiqué dans Pascal. Par exemple, la date peut être plus précise qu'une indication stricte de l'année, sous la forme "1999-10", ou "1999-03-15" ou 1999-03/1999-04...

3.3.3. Le volume du périodique

Les données de ce champ ont dû être reformatées car le volume a une forme variable.
Par exemple, 99-7 (traduit en 0099), v E81-C (traduit en 0081), 9A (traduit en 0009).

---

[16] L'ISSN (ou International Standard Serial Number) est le code international normalisé sur 8 caractères alphanumériques qui permet d'identifier toute publication en série.
[17] Pour plus de détails, voir http://www.isinet.com/journals/
[18] Pour plus de détails, voir http://www.issn.org:8080/French/pub/

Dans le fichier Pascal 1999, 649 notices sur un total de 481 646 (0,13%) n'ont pas pu être converties soit parce que le volume était un mois (nov, ou octnov...) soit parce qu'il était sur plus de 4 caractères (373 notices).

3.3.4. La première page de l'article

Les données de ce champ ont dû être reformatées car la pagination a une forme variable.
Par exemple : A1-A11 (traduit en 0001), 12-15 (traduit en 0012), S.B11-S.B25 (traduit en 0011) ...
Dans le fichier Pascal 1999, 18 150 notices sur un total de 524 596 (3,46%) n'ont pas pu être converties automatiquement ou manuellement car l'information n'était pas présente (cas des monographies principalement).

3.4. Contrôles des appariements après la première fusion

Un premier appariement entre les notices des deux corpus SCI et Pascal a été fait.
Sur ce premier appariement, pour tous les périodiques ayant les mêmes ISSN dans les bases Pascal et SCI, le nombre de notices obtenues lors de la fusion a été comparé à celui des deux corpus initiaux.
Plusieurs cas ont été observés :

- ✓ le nombre de notices est le même (ou très peu différent) dans les 3 corpus SCI, Pascal et corpus fusionné : l'appariement s'est bien déroulé ;
- ✓ le nombre de notices est nul dans le corpus fusionné alors qu'il est identique dans les deux corpus initiaux : l'appariement ne s'est pas fait ;
- ✓ tous les cas intermédiaires.

Nous avons étudié ces cas intermédiaires lorsqu'il y avait une forte disproportion entre les différents corpus (le nombre de notices fusionnées est faible alors que ce nombre est quasiment identique dans les corpus initiaux).

Ensuite, nous avons contrôlé manuellement les notices dans les deux corpus initiaux pour comprendre les raisons du non appariement des notices. Puis, nous avons effectué les corrections manuellement et relancé l'appariement.

Les raisons du non appariement sont de plusieurs ordres et montrent les limites des conventions d'appariement choisies compte tenu de l'hétérogénéité des données des corpus initiaux :

- ✓ Volume du périodique :
  Les volumes doubles ne sont pas toujours indiqués de la même façon dans les 2 bases. Par exemple, indication du volume 0121 dans Pascal et 0120 dans SCI. Cependant cette différence n'est pas systématique. Dans SCI, c'est toujours le premier volume suivi du deuxième, qui est mentionné dans le cas de volumes doubles. Dans Pascal cela peut varier au cours d'une même année (un volume 158-160 est parfois indiqué en 158, parfois 160).
  Volume différent dans les 2 bases : Par exemple pour l'année 1999, pour le "American journal of physiology-endocrinology and metabolism", les volumes 39 et 40 (Pascal) renvoient sur les volumes 276 et 277 (SCI) mais pas pour les autres années.
  Parfois, il faut tenir compte d'informations contenues dans les champs volume et pagination : Pour "Physical review D", il faut ajouter les 2 premiers caractères du champ "pagination" au champ "volume" de Pascal pour obtenir le "volume" de SCI.

- ✓ Pagination de l'article :
  La pagination est cataloguée différemment dans les deux bases. Par exemple, dans SCI la page est 1303, dans Pascal, 041303 (traduit par le programme automatique en 0413).

- ✓ ISSN du périodique :
  Deux périodiques peuvent avoir le même titre (et avoir été associés au même ISSN) mais être différents. Dans ce cas, ils n'ont pas forcément les mêmes volumes et paginations d'articles mais l'appariement a pu se faire quand même sur quelques notices de manière aléatoire. (exemple: la Revue du rhumatisme (ISSN 1169-8446) existe sous la forme d'une version

française et d'une version anglaise. Les articles ne sont pas paginés de la même manière dans les 2 versions).

Nous sommes conscients qu'il subsiste encore des notices non liées bien qu'identiques dans les deux bases, malgré ces différents contrôles visant à réduire leur nombre.
Les résultats obtenus lors de la deuxième fusion sont les suivants :

Tab1 : comparaison du nombre de notices appariées entre la première et la deuxième fusion

Année de Publication	Première fusion	Deuxième fusion	Notices supplémentaires appariées	Pourcentage de notices supplémentaires appariées
1996	328 718	330 881	2 163	0,65 %
1997	349 955	350 390	435	0,12 %
1998	351 454	355 575	4 121	1,16 %
1999	354 122	368 261	14 139	3,84 %
2000	360 716	367 741	7 025	1,91 %

3.4.1.1. Remarques sur les appariements entre les bases SCI et Pascal

Une notice de SCI (identifiable par une clé unique) peut fusionner avec plusieurs notices Pascal différentes et inversement. C'est le cas lorsque deux articles différents ont les quatre mêmes paramètres choisis pour l'appariement. Par exemple, deux très courts articles publiés sur la même page d'un périodique auront le critère "page de début" identique.
Ainsi, retrouve-t-on environ 4 à 5% (selon les années) des notices fusionnées SCI associées à plusieurs notices Pascal, ou une notice Pascal qui a fusionné avec plusieurs notices SCI (pour 1999, cela représente 16 232 notices sur les 354 122 notices qui ont été fusionnées).

4. Résultats

4.1. Sur la formulation des axes du CAP

A titre d'exemple, l'axe "Le vivant et ses enjeux sociaux", a d'abord fait l'objet d'une recherche très précise et limitée. Après avis d'experts du Département des Sciences de la Vie du CNRS, l'ensemble des travaux concernant la biologie et la médecine a été considéré comme pertinent.

4.2. Comparaison des bases de données

4.2.1. Recouvrement des bases

Les pourcentages de recouvrement des deux bases de données SCI et Pascal pour les années de publication 1996 à 2001 ont été étudiés. Ces pourcentages évoluent légèrement selon les années de publication.

Tab2 : nombre de publications dans les différents corpus étudiés et pourcentage de recouvrement

Années de publication	Pascal	SCI	Fusion	Pourcentage Pascal	Pourcentage SCI
1996	436 684	730 314	330 881	75,8	45,3
1997	455 636	754 283	350 390	76,9	46,5
1998	463 782	758 504	355 575	76,7	46,9
1999	466 998	786 632	368 261	78,9	46,8
2000	471 602	802 024	36 7741	78,0	45,9
2001	480 086	792 290	375 971	78,3	47,5

- ✓ Parmi les articles du SCI, environ 46% se trouvent dans Pascal ;
- ✓ Parmi les articles de Pascal, environ 77% des articles se trouvent dans SCI.

4.2.2. Répartition des publications et des périodiques par type de document et par domaine

Afin de mieux connaître les différences et les similitudes entre les deux bases et pouvoir évaluer la qualité des résultats obtenus pour les axes du CAP, la répartition des publications et des périodiques par type de document puis par domaine a été analysée.

4.2.2.1. Par type de document

L'étude s'est faite sur les années de publication 1997 et 1999.
La politique documentaire de la base Pascal n'est pas d'indexer la totalité des informations contenues dans les revues ("cover to cover", couverture intégrale des revues) contrairement à ce qui est pratiqué dans la base SCI. Ainsi, on constate que le corpus résultant de la fusion entre les bases Pascal et SCI, contient presque exclusivement des types de document "article" (plus de 92%) et très peu de type de document "letter" ou "editorial" (environ 1% chacun alors qu'il représente environ 4% dans le corpus initial SCI).
Ainsi, peut-on affiner les pourcentages de recouvrement des bases précédemment calculés qui prenaient en compte l'ensemble des types de document : plus de 58% des types de document "article" catalogués dans SCI se trouvent dans la base de données Pascal alors que le résultat sur l'ensemble des documents est d'environ 45%.

4.2.2.2. Par domaine

L'étude s'est faite sur les années de publication 1997 et 1999.
La classification en huit domaines en usage à l'OST a été utilisée. Chaque périodique est rattaché à un domaine principal, ou à plusieurs pour les périodiques multidisciplinaires (dans ce cas, un coefficient de pondération est attribué).
La répartition des publications en fonction des domaines et le pourcentage de recouvrement entre les deux bases pour 1999 ont été définis pour l'ensemble des documents répertoriés (les résultats pour 1997 sont très voisins de ceux obtenus pour 1999).

Tab3 : recouvrement des bases en fonction des domaines pour tous les types de document (année de publication 1999)

Domaine (classification OST)	Pourcentage de recouvrement Pascal/SCI*
Sciences de l'ingénieur	69,7
Physique	63,3
Sciences de l'univers	54,5
Biologie appliquée	45,8
Recherche médicale	45,0
Chimie	44,1
Biologie fondamentale	31,9
Divers (Mathématiques et périodiques multidisciplinaires)	23,4
Tous domaines	45,2**

* articles du SCI présents dans Pascal
** Le pourcentage de recouvrement "tous domaines" (45,2) (Tab3) est un peu différent de celui du tableau 2 (46,8) car un certain nombre de périodiques ne sont rattachés à aucun domaine et ne sont donc pas comptabilisés dans la répartition par domaine.

Les "Sciences de l'ingénieur" sont bien représentées dans Pascal. Par contre, la "Biologie fondamentale" n'est pas très bien couverte dans Pascal puisque le taux de recouvrement n'est que de 31,9 %.
Ces résultats ont été affinés en étudiant uniquement les références de type "article" (Tab4).

On retrouve alors les mêmes tendances que pour l'ensemble des types de document sauf pour la "Recherche médicale" où le pourcentage de recouvrement devient alors du même ordre que pour les Sciences de l'ingénieur (alors qu'il est très inférieur sur l'ensemble des publications, voir Tab3).

On peut conclure (Tab4), de manière générale, que la majeure partie de l'information scientifique contenue dans les périodiques, dans les différents domaines, se trouve dans des documents de type "articles" (la moyenne est de l'ordre de 70%) mais avec de grandes variations en fonction des domaines. Ainsi, pour la Recherche médicale, environ 50% de l'information se trouve dans d'autres types de document ("meeting abstract", "letter"...), alors que pour la Physique, 94,5% de l'information est dans les articles.

Tab4 : comparaison du pourcentage de type de document "articles" en fonction des domaines dans les différents corpus (année de publication 1999).

Domaines	**SCI**			**Fusion SCI/Pascal**		
	Nombre de publi.	**Nombre d'articles**	**Pourcentage d'articles/publi. Total**	**Nombre de publi.**	**Nombre d'articles**	**Pourcentage d'articles/publi. Total**
Physique	80 378	75 978	94,5	50 902	48 056	94,4
Sciences de l'ingénieur	51 036	43 580	85,4	35 557	33 783	95,0
Chimie	100 762	81 818	81,2	44 426	41 599	93,6
Sciences de l'univers	38 292	34 195	89,3	20 863	19 775	94,8
Biologie appliquée	46 756	37 437	80,1	21 413	20 488	95,7
Biologie fondamentale	128 504	93 848	73,0	40 971	37 756	92,2
Recherche médicale	279 796	150 675	53,9	125 849	109 219	86,8
Divers	56 541	38 419	67,9	13 251	11 662	88,0
Nombre total de publications	782 065	555 950	71,1	353 232	322 338	91,3

4.2.2.3. Comparaison des facteurs d'impact dans les bases de données Pascal et SCI

Nous avons comparé les facteurs d'impact des revues par domaine dans la base SCI et dans la base résultant de la fusion entre les bases SCI et Pascal (cf. Tab5) :

Tab5 : comparaison des facteurs d'impacts des périodiques dans les différents corpus étudiés pour l'année de publication 1999, en fonction des domaines.

Domaines	**SCI**	**Fusion SCI-Pascal**	**Pourcentage Fusion SCI-Pascal/SCI**
Physique	1,658	1,769	106,7
Sciences de l'ingénieur	0,844	0,838	99,3
Chimie	1,746	1,532	87,7
Sciences de l'univers	1,422	1,378	96,9
Biologie appliquée	1,242	1,314	105,8
Biologie fondamentale	3,168	2,875	90,8
Recherche médicale	2,257	2,303	102,0
Divers	1,468	1,645	112,1

- ✓ Dans les domaines Physique, Biologie appliquée et Divers (Mathématiques et périodiques multidisciplinaires), les périodiques communs à Pascal et SCI ont un meilleur facteur d'impact moyen que dans le corpus SCI seul.
- ✓ Pour les domaines Chimie, Biologie fondamentale et Sciences de l'univers, au contraire, les périodiques communs à Pascal et au SCI ont un facteur d'impact moins important que dans le corpus SCI.

- ✓ Pour les autres domaines (Sciences de l'ingénieur, Recherche médicale), les facteurs d'impact des revues sont quasiment identiques dans les deux corpus.

Ces résultats (Tab5 et des résultats complémentaires non présentés ici) nous ont amené à exclure de l'étude des axes du contrat d'action pluriannuel CNRS-ETAT, l'axe "Astroparticules". En effet, ce domaine est mal représenté dans le corpus fusionné. Les résultats auraient été difficilement interprétables.

5. CONCLUSION

La mise au point d'un réservoir de données issue de la fusion des deux grandes bases de données bibliographiques SCI et Pascal offre des perspectives d'études intéressantes, en permettant d'obtenir des informations assez complètes sur les publications des laboratoires du CNRS en fonction de thématiques précises.
Comme pour toutes les études bibliométriques, il est important de connaître les limites :

- ✓ intrinsèques des corpus à fusionner, liées aux règles de catalogage, à leur couverture géographique...
- ✓ de la méthodologie de fusion utilisée : choix des critères d'appariement, et formatage de ces critères ;
- ✓ de la traduction d'un texte "scientifique" en une requête documentaire,
- ✓ propre au corpus-résultat de la fusion. Ainsi, ce corpus ne contient pas l'ensemble des publications répertoriées dans chacune des deux bases de données.

Aussi, les résultats obtenus devront-ils être validés par des experts du domaine étudié et, le cas échéant, complétés par d'autres études en utilisant par exemple d'autres bases de données et sources d'information.
Cependant, ces études peuvent permettre de donner des tendances de l'évolution de la recherche et d'effectuer des comparaisons au niveau national ou international.

Cette étude bibliométrique pourrait aussi être un bon outil pour améliorer la couverture des périodiques traités dans la base de données Pascal.
En effet, il peut être intéressant pour décider du choix des périodiques retenus pour figurer dans la base Pascal, de renforcer la prise en compte de certains paramètres, comme les facteurs d'impact des revues publiés chaque année par l'ISI, corrélés aux publications des laboratoires CNRS.
Il serait intéressant de comparer régulièrement les périodiques catalogués dans les bases Pascal et SCI afin de compléter ou modifier la couverture des périodiques de la base Pascal, par exemple analyser plus de périodiques dans certains domaines peu couverts comme la Biologie fondamentale, la Chimie ou les sciences de l'univers.

Pascal pourrait ainsi devenir une véritable source de données, en complément du SCI, pour réaliser des études bibliométriques sur les publications scientifiques des laboratoires CNRS.

Note : Certaines données utilisées dans cet article sont tirées de la base Science Citation Index (SCI).

6 Bibliographie

[1] BARRE R., LAVILLE F., TEIXEIRA N. et ZITT M., L'Observatoire des sciences et des techniques : activités - définition - méthodologie, Les sciences de l'information - Bibliométrie, scientométrie, infométrie, sous la direction de Jean-Max Noyer, Presses Universitaires de Rennes, 1995, 219-245.

[2] Comité National d'Evaluation de la Recherche (CNER), Evaluation de la recherche publique dans les établissements publics français, La documentation française, 209 p, 2003.

[3] GAUTHIER E., L'analyse bibliométrique de la recherche scientifique et technologique : guide méthodologique d'utilisation et d'interprétation, 1998, Observatoire des Sciences et Technologies (CIRST), http://www.veilledulendemain.com/fichiers/Etudebiblio.pdf (consulté le 10 septembre 2004).

[4] GRIVEL L., FAGHERAZZI H., FOURNERET P. et ZEROUKI A., La conception de bases de données infométriques : analyse de la pratique de trois observatoires européens et proposition d'une méthode d'intégration de données hétérogènes, Ile Rousse, 27 septembre-1er octobre 1999.

[5] JAGODZINSKI-SIGOGNEAU M., BAUIN S, COURTIAL JP et FEILLET H., Scientific innovation bibliographical databases : a comparative study of the Science Citation Index and the Pascal database, Scientometrics, 1991, 22(1), 65-82

[6] MOED HF., DE-BRUIN RE. et VAN-LEEUWEN TN, New bibliometric tools for the assessment of national research performance : database description, overview of indicators and first applications, Scientometrics, 1995, 33(3), 381-422

[7] RICHARD N. Comparaison des bases SCI et Pascal, Rapport DESS de Gestion du Patrimoine Immatériel de l'Entreprise, Université de Marne la Vallée, 1998.

[8] Unité d'Indicateurs de Politique Scientifique (UNIPS), Les publications des laboratoires CNRS et leur impact (sciences de la matière et de la vie) 1991-2000, 2002, http://www.cnrs.fr/DEP/doc/bib2002.pdf (consulté le 10 septembre 2004).

[9] VIDAL S., DUCLOY J. et HOUDRY P., Mining medical data using multiple corpora interaction : the transcriptomics investigation server experiment, 2003, http://dilib.inist.fr/documents/2003/sci/SCI_2003_S113GI.pdf (consulté le 10 septembre 2004).