



Projet gestion des connaissances et des savoirs. Rapport final

Michel Zitt, Elise Bassecoulard

► To cite this version:

Michel Zitt, Elise Bassecoulard. Projet gestion des connaissances et des savoirs. Rapport final. 2004.
hal-02828056

HAL Id: hal-02828056

<https://hal.inrae.fr/hal-02828056>

Preprint submitted on 7 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PROJET GESTION DES CONNAISSANCES ET DES SAVOIRS

Sous-groupe projet bibliométrie¹

Rapport final– 10/02/04 Michel Zitt (chef de projet)

Destiné aux membres INRA du groupe

Pour discussion avant version finale diffusable

Participants extérieurs

Lesya Baudoin (INSERM), Serge Bauin (CNRS), Arnaud Muret (CEA), Nicole Pinhas (INSERM), Françoise Rousseau-Hans (CEA), Jacques Souillot (CNRS), Milorad Stjepanovic (IRD)

Participants INRA (structures 2003)

Chef de projet Michel Zitt (ESR-Nantes, OST-PARIS)

Emmanuelle Alfonsi (DISI - Corse), Elise Bassecoulard (ESR-Nantes), Marie-Hélène Cathala (DISI-Angers) Annie Chartier (DISI-Expertise), Veronique Garandel (Réseau doc SAD), Anne-Marie Gogüe (DS SED), Geneviève Lacombe (DISI), Hugues Leiser (EAG-Avignon) Marie-Hélène Magri (DISI - Jouy en Josas), Hubert Pampouille (DISI SI), Christine Sireyjol (DISI SI)

Les animateurs remercient les participants de leurs contributions, en particulier pour les présentations de produits et les suggestions pour les documents de travail (fiches bibliométriques et questionnaire).

INTRODUCTION

L'exploitation de l'IST par des méthodes quantitatives (bibliométrie) répond à des besoins grandissants de divers utilisateurs : ingénieurs documentalistes, chercheurs, départements et centres, direction générale. Selon les utilisateurs, la forme de cette demande varie : organisation de l'information documentaire, veille scientifique, indicateurs stratégiques de publication. Le but du groupe de travail est de préciser les contours de cette demande et de l'offre correspondantes, autant en ressources humaines (pôles de compétence) que matérielles (méthodes et instruments), afin d'éclairer et de coordonner les réponses et de promouvoir, si possible, des solutions mutualisables. Ceci est d'autant plus nécessaire que la pression de la demande risque de faire éclore des solutions mutuellement incompatibles et, parfois, déconnectées de "l'état de l'art" bibliométrique.

Il y a de multiples expériences sur lesquelles s'appuyer, à l'intérieur et à l'extérieur de l'organisme.

- l'écoute de la demande à l'intérieur et à l'extérieur de l'INRA.
- la recherche de solutions pouvant faire l'objet d'une capitalisation et d'une mutualisation.
- la coordination avec les autres groupes travaillant sur des sujets liés : système d'information, bases de données, nouveaux métiers, nouvelles compétences, gestion des connaissances quand il sera mis en place.

Trois finalités de la bibliométrie dans un institut de recherche

On peut distinguer trois grandes finalités dans la pratique bibliométrique des institutions. Chaque finalité correspond à des types d'analyses avec des contraintes particulières.

- le suivi de la dimension "publications" dans la production de l'institution (optique tableau de bord)

¹ Groupe de travail mis en place à la demande de Geneviève Lacombe, directrice déléguée DISI-IST, dans le cadre de la politique IST axe 3 : « L'organisation et la gestion des savoirs : une réponse à construire aux demandes sociales d'expertise et de veille scientifique ».

- le positionnement stratégique des activités de recherche
- (pour mémoire) la veille scientifique ou technologique

Les applications de la bibliométrie ne se bornent pas, bien entendu, aux trois finalités décrites ci-dessus. Il y notamment un nombre croissant d'applications de la bibliométrie en économie de la science et du changement technique, et en études de politique scientifique (ne serait-ce que l'analyse du lien science-technique). Nous nous sommes limités ici aux applications opérationnelles.

I- LE SUIVI DE LA DIMENSION "PUBLICATIONS" DANS LA PRODUCTION DE L'INSTITUTION.

11. Contexte général:

La "rose des vents" de la recherche (CSI-Mines) décrit les multiples dimensions de l'activité des laboratoires, selon la destination des productions (le monde académique, l'industrie, la puissance publique, le grand public). La dimension de la "recherche académique" n'est qu'une de ces dimensions, mais naturellement elle est essentielle dans les EPST. Les publications autres qu'académiques relèvent des méthodes bibliométriques à transposer au cas par cas :

- CR de conférences
- littérature de transfert
- littérature grise, supports de formation etc
- brevets et autres titres de propriété intellectuelle
- rapports pour l'aide à la décision publique ou privée, rapports d'expertise
- compte-rendus de mission
- documents de relations contractuelles
- etc.

La transposition des méthodes bibliométriques est d'autant plus aisée que les documents d'outputs sont codifiés.

Les indicateurs de publication au sens large sont un volet majeur du dispositif d'indicateurs requis dans les contrats pluriannuels engageant les institutions, dispositifs à renforcer dans le cadre de la LOLF. Ils sont tout aussi utiles dans une optique de gestion scientifique interne. Tous les niveaux de décision de l'institution sont potentiellement concernés, en particulier la direction, les départements scientifiques et les centres.

Bien que partageant avec les processus d'évaluation collective et individuelle la nécessité d'un suivi des publications, la bibliométrie n'a pas vocation à se substituer à la démarche synthétique qui caractérise les processus d'évaluation collective et a fortiori individuelle (et qui sont encadrés par la loi). L'utilisation de la bibliométrie en évaluation, en particulier pour de petites entités, ne peut se faire qu'avec une grande prudence, la finalité de la bibliométrie étant d'abord de fournir des indicateurs statistiques.

En tenant compte des sources de données disponibles, on peut distinguer deux types de tableaux de bord selon leur objectif :

- le suivi de l'ensemble des productions de l'institut, où l'on cherche l'exhaustivité, avec codification des types de sortie.
- le tableau de bord avec valeurs de référence (p.ex. production française, européenne, mondiale), forcément limité à un type de source pour lequel la comparaison est possible (en pratique les sources ISI pour les publications scientifiques), les bases de brevet des systèmes français, européen, américain (+ PCT).

12 données d'entrée

La production d'indicateurs bibliométriques résulte d'un compromis entre la disponibilité des données de publication et les besoins de l'organisme en matière de gestion ou d'analyse stratégique. Dans la grande majorité des cas, les sources internes (systèmes documentaires) ou externes (base de données bibliographiques, sources hétérogènes comme le Web), n'ont pas été conçues pour la finalité

"indicateurs", ce qui impose une vigilance particulière. De toute évidence, s'agissant des sources internes, un rééquilibrage des cahiers des charges en faveur de la finalité "indicateurs" est nécessaire.

Les analyses bibliométriques sont plus aisées à partir de données codifiées et documentées, des données hétérogènes limitant souvent à des exploitations "*quick and dirty*", parfois acceptables dans une optique de veille mais certainement pas dans une optique "tableau de bord", et encore moins pour la production d'indicateurs « certifiés » destinés aux autorités de tutelle.

Cette nécessité de codification et de documentation s'applique aux données quelle que soit leur origine, interne ou externe (voir Annexe I).

Les contraintes pour une utilisation des données à des fins bibliométriques :

- la documentation précise des données utilisables, de leur provenance et de leurs caractéristiques
- pour la partie dépendant du système interne, le respect des bonnes pratiques documentaires (dédoublonnage évidemment, contenu des champs) est une exigence minimale mais non suffisante. La finalité indicateurs exige par exemple la saisie détaillée de l'ensemble des adresses d'auteurs.
- contrainte spécifique à la finalité "tableau de bord", la nécessité d'une couverture large, couvrant les productions de l'institut non répertoriées dans des bases externes aisément accessibles. Par conséquent, un système quasi-exhaustif de couverture des productions de l'institut est à recommander, avec des mécanismes facilitant et garantissant la saisie.
- la combinaison d'informations existant dans diverses bases sur les publications est souhaitable dans la limite du juridiquement possible. Exemple: citations du SCI, codes de classement PASCAL, index de CAS ou Medline etc. .

Les indicateurs de tableau de bord nécessitent au minimum un **repérage fiable des publications** de l'institut dans leurs différentes formes.

- pour les tableaux de bord avec référence externe (France, Europe, Monde), le repérage doit s'effectuer dans les bases externes fournissant ces données de cadrage, en pratique les données ISI. Dans ce repérage², la base interne peut être utilisée avec profit comme "germe", voire comme support de « revendication ».
- pour des tableaux de bord plus exhaustifs, la combinaison de sources internes et externes est recommandable.

Le repérage des **brevets et COV** devrait être en principe facilité par l'exigence d'exhaustivité pour le suivi du portefeuille de propriété intellectuelle (DISI, INRA transfert SA, Agri-Obtentions...).

Les indicateurs bibliométriques, s'agissant des outputs, mettent en jeu des sources d'origine diverse. Par ailleurs, des mesures d'input entrent dans les indicateurs mixtes, p.ex. input-output, qui font appel aux **autres volets du système d'information** (ressources financières, ressources humaines, valorisation). La nécessité de traiter ces indicateurs d'une manière cohérente prêche pour des infrastructures informatiques inter-opérables, démarche récemment réaffirmée pour les enjeux du S2I (INTERCOM 91 23/07/03 ³).

13 indicateurs typiques et leurs contraintes

Le produit typique, dans le cas de publications "académiques", est le tableau de bord sur grille prédéfinie (thématique scientifique, objectifs socio-économiques, découpage administratif...) avec comme principaux volets :

- la production : volume et spectre des publications. Cette production ne peut être rapportée aux ressources humaines ou financières (données de productivité) qu'avec les plus extrêmes précautions.

² exemple, le repérage effectué jusqu'en 2003 pour l'INRA par le LERECO-ESR Nantes pour la coopérative IPSIR

³ articulation autour de « noyaux d'information commun d'espaces de liberté propres à chaque niveau d'organisation »

- la collaboration scientifique : volume, spectre et affinités de co-publications ; d'autres relations sont accessibles, comme les interdépendances scientifiques à partir des réseaux de citations.
- la visibilité à partir des citations, si possible avec référence nationale ou internationale.

Le détail des indicateurs classiques figure en Annexe II.

Les contraintes de l'approche "tableau de bord" sont les suivantes :

- codification et standardisation des entrées
- approche quasi-"comptable" et pérenne
- selon la finalité, exhaustivité ou comparabilité.

14 méthodes et outils

Par contraste avec les analyses de positionnement (finalité 2), les tableaux de bord n'exigent pas en général de traitements de grands ensembles de données, les références nationales ou internationales pouvant se limiter à des données agrégées.

Il n'y a pas de spécificité bibliométrique forte sur les outils. Il y a une spécificité des objets, qui demande certaines adaptations ou paramétrages des méthodes, mais en gros l'ensemble de techniques de stockage, manipulation et traitement statistique de données est mobilisable pour les études bibliométriques :

- stockage et interrogation : bases de données documentaires, bases de données relationnelles, logiciels généraux (SAS) intégrant des fonctionnalités de SGBD, outils de requêtage et "reporting" (BI, BO, SAS Entreprise Guide), moteurs de recherche sur bases hétérogènes.

- traitement et visualisation des données : l'exploitation des données de tableau de bord nécessite en général des outils assez simples de tabulation et de graphique. Pour des exploitations ou visualisation plus élaborées, les outils généraux de traitement statistique et d'analyse de données et de cartographie peuvent être utilisés.

Pour la finalité "tableau de bord", des outils simples de base de données sont en général suffisants, allant du plus léger (Excel), au plus lourd (Oracle, SAS) en passant par les intermédiaires. L'harmonisation des données entre départements et au niveau global plaide pour une solution commune. Notons que des outils intégrés (Entreprise Guide de SAS, Clementine SPSS) offrent une solution conciliant interactivité et puissance. Quelques indications complémentaires figurent en annexe III.

15 état actuel à l'INRA

On distingue actuellement diverses méthodes pour le suivi des publications/productions, au niveau global et par département :

- une base des publications, PUBINRA, conçue initialement à des fins documentaires et d'affichage externe, alimentée par le réseau documentaire. La refonte de PUBINRA est en cours (DISI-IST, projet PRODINRA, C. Sireyjol)
- dans le cadre de la coopérative d'indicateurs des institutions publiques (groupe Ministère de la Recherche, présidé par J. Sevin), un repérage des publications de l'INRA dans la base internationale de référence est fourni (actuellement par le LERECO-ESR) à l'opérateur technique, l'OST.
- les indicateurs d'institution collectés dans le cadre des directoriales par la DIFAG (F. Sevin J. Dodet) comportent des chiffres agrégés de publication déclinés par département et centre, avec une ventilation thématique pour le sous-ensemble affectable à une spécialité ISI.

Cette tâche est actuellement effectuée par le LERECO-ESR à partir du repérage ci-dessus et de PUBINRA.

- plusieurs départements ont leur propre système de suivi des publications

Certains tableaux de bord peuvent être gérés en interne, mais pour intégrer les référentiels nationaux et internationaux d'une part, les données de visibilité d'autre part, une sous-traitance à un partenaire disposant de ces données peut être envisagée (voir plus loin les opportunités du côté OST).

Le chantier en cours sur la refonte des systèmes d'information inclut la réflexion sur le dispositif d'alimentation, notamment la répartition des tâches et la coordination entre les différents niveaux institutionnels (unité, départements, centres...) et les modalités d'exploitation de ces systèmes.

La demande institutionnelle est de plus en plus forte et devrait perdurer, notamment au niveau des départements, pour accompagner les réflexions sur les refontes de périmètre, les évaluations périodiques, les schémas stratégiques, etc... Les actions de sensibilisation à la bibliométrie (voir IV, recommandations) doivent donc toucher les personnes concernées hors DISI dans les départements et centres.

II - POSITIONNEMENT ET EVALUATION STRATEGIQUE.

21 Contexte

Le besoin de positionnement "stratégique" des acteurs dans leurs domaines d'action présente ou future, avec une volonté comparative et prospective, est partagé par les acteurs économiques et les chercheurs. La bibliométrie offre une "boîte à méthodes", reposant elle-même sur les boîtes à outils des différentes branches d'analyse quantitative. Nous partons du principe que, pour justifier d'une analyse bibliométrique, la question étudiée doit être d'ordre statistique, et mettre en jeu un grand nombre d'événements. Si le positionnement dépend d'événements isolés (manquer ou non l'information sur la sortie d'un brevet fondamental), on se retrouve dans une problématique de "veille" (objectif 3). L'exhaustivité n'est pas en général requise dès lors que les données respectent un principe de sélection des meilleures sources - ce qui est en général le cas des bonnes bases de données, une fois validé le périmètre.

Questions typiques relevant d'analyse de positionnement :

- caractérisation stratégique des institutions concurrentes et/ou partenaires :
- réseaux de partenariats et/ou de co-activité
- sur bases scientifiques ou sur un ensemble S&T comprenant les activités de propriété intellectuelle.

22 Données

La source externe de référence (ISI) est essentielle pour le positionnement et l'évaluation stratégique. L'avantage des sources ISI est l'accès large aux données de collaboration et de citation, qui compense l'absence de plan de classement et de descripteurs d'indexation. La seule alternative multidisciplinaire est Pascal (Inist), dont la couverture et la stabilité laissent à désirer, mais qui a l'avantage de comporter des descripteurs d'indexation et un plan de classement.

Pour des analyses thématiques, selon les domaines, la couverture des bases multidisciplinaires peut s'avérer biaisée ou insuffisante (voir Annexe I). Le recours à des bases spécialisées, notamment les grandes bases thématiques est alors nécessaire en substitution ou complément des bases ISI. Chacune de ces bases a ses propres caractéristiques, avec en général des index très performants mais une saisie des affiliations insuffisante.

Du côté brevets, chaque système a sa propre base (EPAT pour l'OEB, USPTO pour l'USPO, etc.). Derwent/WPI est multi-systèmes, avec une forte valeur ajoutée dans les traitements.

En tant que de besoin, d'autres sources peuvent bien entendu être mobilisées, y compris sur Internet. L'analyse bibliométrique est naturellement plus délicate sur une information peu codifiée et hétérogène.

Le contenu des champs (après combinaison éventuelle de sources) détermine la base des méthodes bibliométriques utilisables. Par exemple, pour l'analyse des thématiques d'un domaine, des sources à indexation excellente se prêtent particulièrement bien aux méthodes lexicologiques. Les méthodes de citation ne seront applicables qu'au périmètre couvert par les bases ISI ou par des systèmes alternatifs (CiteSeer en informatique au sens large).

23 Indicateurs typiques

Les plans de classement des bases de données, aussi détaillés et pertinents qu'ils puissent être, gèrent difficilement la dynamique de recomposition permanente qui caractérise l'activité scientifique. Les analyses de positionnement concurrentiel d'un acteur supposent typiquement :

- la délimitation du domaine, qui peut s'avérer extrêmement délicate notamment pour les champs transversaux ou en voie d'émergence ; cette phase conditionne largement la qualité des résultats
- la structuration du domaine à étudier en un ensemble de thèmes ou fronts de recherche en interaction.

Pour délimiter et structurer le domaine :

- on sélectionne une ou plusieurs relations structurantes, par exemple la parenté lexicale (mots associés), citationniste (couplage et co-citation), l'association de codes de classement, plus rarement les liens de co-publication.
- et des techniques statistiques et d'analyse de données appropriées à la discrimination des thématiques sur la ou les relations structurantes.

Les séquences d'analyse "micro-bibliométrique" d'un champ combinent typiquement des phases d'analyse lexicométriques et/ou citationnistes, en sollicitant les commanditaires et les experts du domaine aux phases clés (voir ***Annexe II-3 le détail des opérations).

Les combinaisons de méthode sont particulièrement utiles lors des étapes de délimitation du domaine, où les techniques documentaires traditionnelles, les méthodes lexicales et les techniques de citation peuvent être mobilisés sur leurs points forts, elles sont aussi envisageables pour l'identification des thématiques de recherche et de leurs liens.

L'analyse des réseaux permet de mettre en évidence les propriétés des individus ou des groupements identifiés, étudiées par la théorie des graphes et plus récemment par la théorie des réseaux sociaux : densité, centralité, équivalences, passerelles, chemins et propriétés globales du graphe (voir Annexe II). La combinaison de caractéristiques structurelles et dynamiques permet de proposer, avec prudence, l'identification de thèmes en voie d'émergence.

Sur chaque thème identifié, l'analyse reprend typiquement les indicateurs classiques sur les acteurs et leurs relations, mentionnés précédemment : volume et spectre ST d'activité ; visibilité ; relations dans le réseau : co-publications, relations de citation, etc. Sous réserve d'une identification correcte des acteurs, laboratoires ou individus, des interprétations en termes de ressources humaines, par exemple la sélection de partenaires, sont possibles.

Le positionnement suppose une comparaison avec les partenaires/ concurrents internationaux, et donc des traitements plus élaborés et coûteux qu'un tableau de bord où on se limite à indiquer des références nationales/internationales très agrégées.

Les analyses de positionnement valorisent en général :

- la sélectivité (plutôt que l'exhaustivité)
- la robustesse (plutôt que l'immédiateté)

Les capacités des méthodes d'analyse de réseau (de publications, d'acteurs, de liens linguistiques ou de citations) sont également très précieuses pour aider à *la mise en place de référentiels thématiques* (détection de périmètres de disciplines, de recouvrements, etc.) qui ont vocation à être utilisés non seulement pour le positionnement thématique, mais aussi le cas échéant pour les tableaux de bord ⁴.

24 outils

L'ensemble des outils de traitement de l'information est mobilisable pour l'analyse stratégique (voir Annexe II).

- outils spécialisés intervenant à un stade d'étude, ou présentant une méthode d'analyse particulière, ex. moteurs, programmes d'analyse lexicale, d'analyse statistique spécialisée sur quelques méthodes (Le Sphinx), etc.
- logiciels statistiques multi-fonctions, avec orientation bibliométrique (ex. Tétralogie)
- offres globales plus complètes, du type logiciel statistique général avec fonctions de fouille de données (ex. SAS- Enterprise Miner, SPSS- Clementine, IBM- Intelligent Miner).

25 état actuel à l'INRA

Contrairement aux tableaux de bord qui sont une activité relativement facile à cadrer, à la fois pour le cahier des charges et les outils, les analyses de positionnement sont un domaine foisonnant, pouvant mobiliser des techniques d'analyse de données, de graphe et de cartographie très variées. C'est aussi un domaine où la prestation de service se développe, appuyée ou non sur des logiciels propres (exemple: Tétralogie). A l'INRA (et semble-t-il dans les autres organismes) c'est une activité émergente et peu structurée, avec des initiatives au coup par coup⁵ de la part des départements de recherche ou de la direction.

III - LA VEILLE SCIENTIFIQUE ET TECHNOLOGIQUE

Nous rappelons pour mémoire le domaine de la veille⁶, qui est traité par un autre groupe de travail (voir la liste de diffusion mise en place par le Pôle Partenariat Transfert Innovation de J-F Quillien). Pour l'essentiel la veille est une activité jusqu'à présent décentralisée et prise en charge par chaque chercheur ou laboratoire. A grande échelle, la veille peut être considérée comme faisant partie de la prospective (DADP, puis DS SED).

31 Contexte

La veille technologique-scientifique a pour objet la détection d'événements ou de tendances significatives pour l'acteur, venant compléter son information propre sur son aire de compétence spécialisée. L'exploitation des données, soit dans les bases spécialisées soit sur Internet, n'est qu'une

⁴ c'est le cas par exemple des « sous-disciplines » INRA/OST

⁵ des expériences sur Tétralogie et Leximine ont été présentées lors de la journée pour les Départements de recherche de décembre 2002, à l'initiative de la DISI-IST (G. Lacombe)

⁶ qui comporte aussi des dimensions économique, juridique, commerciale ...

partie de l'activité de veille, qui comporte de multiples aspects qualitatifs ("intelligence économique", etc.).

On peut distinguer deux modalités :

- la recherche d'information ou d'"indications" à l'échelle micro : éléments d'information pertinente à dégager d'ensembles massifs et hétérogènes (recherche de "pépites dans la gangue", notamment dans la gangue Internet).

Exemple : un engagement d'investissement d'un concurrent, une mobilité individuelle d'un chercheur-clé, une recherche de recrutement ou de partenariat, un dépôt de brevet bloquant, etc... L'immédiateté de l'information est souvent critique, l'interaction avec la demande est forte, et la connaissance du domaine importante pour rejeter l'ivraie.

- la veille à l'échelle "macro", p.ex. le dynamisme comparé des domaines de recherche avec identification de tendances émergentes.

Cet objectif est partagé avec les analyses de positionnement stratégique ci-dessus.

32 Données

La veille n'a guère d'exclusive sur les données à mobiliser (enquêtes, Internet, bases de données structurées, etc.) et explore des données massives, si possible exhaustives, dont le tri peut faire appel à des méthodes statistiques sophistiquées intégrées aux outils de fouilles de données (cf infra). Toutefois la mise en oeuvre « à l'aveuglette » de gros outils de data-mining sur sources hétérogènes n'est pas forcément la panacée. L'exploitation systématique de bases de données structurées par des spécialistes du domaine rompus aux techniques documentaires reste aussi efficace dans certains contextes.

33 informations et indicateurs typiques

Ils associent en général les aspects thèmes et acteurs :

- domaines/ acteurs concurrentiels, domaines/ acteurs en voie d'émergence
- détection de compétences, détection de partenariats possibles
- description et cartographie d'un environnement chercheur : réseau de citations, etc...

34 méthodes et outils

La "fouille de données" (data-mining) de l'ère Internet associe généralement des moteurs de capture d'information et de navigation qui s'ajoutent au requêtage classique de bases de données, des outils d'analyse de données ou de graphes reconfigurés pour l'analyse de fichiers massifs et hétérogènes, l'exécution en série ou en parallèle d'analyses statistiques, avec des ergonomies accessibles au non spécialiste. Les outils sont sensiblement les mêmes que pour le positionnement stratégique, des plus spécialisés (moteurs classiques de l'Internet, outils statistiques spécialisés p.ex. modules d'analyse factorielle, outils cartographiques) aux offres plus globales de fouille de données et d'analyse signalés plus haut.

Caractéristiques typiques de la fouille de données :

- traitement massif de données hétérogènes, avec parfois un objectif d'exhaustivité
- combinatoire des méthodes
- souci d'immédiateté

IV - CONCLUSION : RECOMMANDATIONS

Les recommandations sont déclinées en deux volets. Le premier récapitule les actions sur les objectifs précédents, le second aborde les actions plus transversales. Le message essentiel est que les indicateurs bibliométriques ne relèvent pas au premier chef de l'activité documentaire mais de l'analyse stratégique. Ils peuvent en revanche bénéficier grandement du réseau de compétences documentaires et bibliométriques de l'institut.

41. par objectif

préalable : la question des données

- réaffirmer la complémentarité des sources externes (pour la production académique) et internes (tous types d'information)

- introduire les critères d'usage bibliométrique dans la conception des systèmes d'information (voir ***Annexe I sur les sources internes)

Exemples :

- + permettre l'identification des acteurs, p.ex. des affiliations étrangères, sur les productions (PUBINRA-EX LIBRIS-PRODINRA)
- + faciliter les structurations thématiques etc.

- prévoir l'interfaçage avec des référentiels internes ou externes accessibles.

la finalité tableau de bord

- la production de **tableaux de bord pour le suivi d'activité**, normalement ancrés sur une base interne repérant l'ensemble des outputs, peut faire l'objet d'un "service bibliométrique minimum" (production/spécialisation, collaborations, visibilité espérée) qui peut être implémenté sans grande difficulté quelle que soit la plate-forme de traitement. La structure en base de données (Access, Oracle, etc...) de préférence couplée ou intégrée à un système statistique (comme dans SAS) est plus favorable au calcul d'indicateurs. Il est très recommandable d'utiliser un seul type (ou un nombre restreint) de plate-forme et de protocole, pour garantir la comparabilité des résultats, leur mutualisation possible, et la consolidation au niveau de l'institut.

Ces tableaux de bord et les indicateurs liés peuvent être gérés en interne (assistants CD, PDC, etc.), avec l'appui des compétences bibliométriques présentes à l'INRA, qui peuvent être sollicitées lors de la mise en place des systèmes et des protocoles de calcul.

- pour les **tableaux de bord avec référentiels nationaux et internationaux** d'une part, les **données de visibilité réelle** des publications d'autre part, le recours à un organisme disposant de ces données est à envisager. L'OST, GIP dont l'INRA est membre fondateur, prépare une offre aux institutions de recherche dans ce sens. Sur la base d'accords bilatéraux, les indicateurs de base (production, collaboration, batterie d'indicateurs d'impact) pourraient être périodiquement fournis sur des découpages propres à chaque institution dans la base ISI-Web of Science, dont une version bibliométrique aisément exploitable sera mise en place prochainement à l'OST, avec le soutien des organismes de recherche.

la finalité positionnement stratégique

Les besoins en positionnement stratégique sont déjà ressentis dans les départements et au niveau central, et ils iront en croissant avec la gestion par programmes, la participation grandissante à des GDR, le développement de l'activité en réseau, et indirectement la mise en place de la LOLF.

L'éventail des choix est ici très large :

- pour les méthodes, des plus légères (analyse factorielle ou classifications sur lexique de quelques centaines de notices) aux plus ambitieuses (structuration de domaines en appui à la prospective)
- pour les modalités institutionnelles qui peuvent privilégier le traitement en interne ou la sous-traitance, ou associer les deux en fonction des opérations prévues et des compétences disponibles.

Une grande partie des demandes peut être instruite par des ingénieurs documentalistes formés à la bibliométrie exploratoire et disposant d'outils appropriés (base de données, logiciel d'analyse statistique légère comme Sphinx). La formation de ces ingénieurs doit également leur permettre d'instruire des sous-traitances externes avec l'aide éventuelle d'un réseau d'experts.

A plus grande échelle, des opérations entre veille et analyse stratégique nécessitent une compétence bibliométrique plus pointue et une bonne capacité de dialogue avec les commanditaires, soit pour réaliser les analyses, soit pour encadrer l'externalisation. Dans les deux cas, l'activation du réseau interne d'experts bibliomètres est souhaitable, afin de garantir les bonnes pratiques en accord avec l'état de l'art. L'expertise bibliométrique est répartie dans diverses branches de l'institut : DISI, ESR-SAE2, MIA (ex BIA) et divers départements scientifiques. Un schéma typique d'analyse thématique est indiqué en Annexe II-3

La mise en place éventuelle de cellules opérationnelles pouvant réaliser en interne des analyses complètes de ce type suppose une professionnalisation et un accès à des plates-formes puissantes de statistique et de fouille de données, type SPSS-Clementine ou SAS-Entreprise Miner. Divers outils spécialisés disponibles en interne, basés ou non sur ces plates-formes, peuvent naturellement être mobilisés en complément.

Si les finalités sont différentes, les tableaux de bord, le positionnement stratégique et la veille partagent pour partie des méthodes et des outils. De ce point de vue, l'accès aux plates-formes multifonctionnelles précitées constituerait un atout. Ces outils sont également utilisables dans d'autres contextes (pex fouille de données pour la bio-informatique).

42 mesures transversales

mesures de sensibilisation

La sensibilisation et la formation des personnels appelés à recueillir et à traiter/orienter les demandes est une priorité. Pour les ingénieurs documentalistes, l'analyse bibliométrique est une des perspectives d'évolution, parmi d'autres dont la communication (voir Groupe Nouveaux métiers, Nouvelles compétences, C. Pétrissant). La sensibilisation comporte divers volets :

- sensibilisation aux méthodes et aux outils (aussi bien des chercheurs que des ingénieurs documentalistes). Des sessions de formation associant le groupe bibliométrie ont été organisées et (mars et décembre 2003, voir Annexe IV) devraient être poursuivies.
- mise à disposition de manuels/ fiches de référence : des "fiches bibliométriques" décrivant les principaux indicateurs et les bonnes pratiques sont en cours d'élaboration. Des fiches prototypes ont été présentées lors des séminaires DISI (voir ***Annexe IV) et un système hypertexte sera monté sur Intranet.

Le développement d'une "culture d'indicateurs" est également indispensable pour prévenir les interprétations abusives des données bibliométriques ("bibliométrie sauvage"), sans doute inéluctables avec la montée en puissance de la demande.

maintenir et développer un réseau de compétences

La compétence bibliométrique est répartie dans divers départements de l'institut. Le réseau de compétence, partiellement mis en place à la faveur des initiatives DISI, est à maintenir et élargir du côté des départements qui sont également des centres de compétences. Ce réseau est le cadre naturel de diffusion de l'information et de discussion des initiatives relatives à la formation, à la coordination, à l'instruction des demandes (remontée des demandes excédant les capacités de traitement locales), d'évaluation de méthodes et d'outils. L'Intranet est un cadre naturel pour l'expression de ce réseau, dans ses activités de formation (fiches bibliométriques), le partage des remontées d'expérience, etc.

Qu'une responsabilité particulière soit confiée aux têtes de réseau pour faciliter les contacts et aiguiller les gros dossiers (réalisation en interne, prestation externes) est une question ouverte.

La question d'une cellule opérationnelle, capable de traiter les grands projets ou d'en assurer un suivi rigoureux, se pose également.

Enfin, le soutien aux équipes actives en recherche et développement directement ou indirectement bibliométrique, pouvant transmettre et nourrir l'état de l'art (SAE2, DISI, MIA), est nécessaire.

manifester le caractère stratégique des indicateurs bibliométriques

L'affichage du caractère stratégique des indicateurs bibliométriques dans l'arsenal des indicateurs d'organisme est une condition nécessaire à des évolutions positives. Cela signifie concrètement :

- que les exigences en matière d'indicateurs soient relayées dans l'institut, en particulier lors de la conception des systèmes d'information,
- et en particulier que l'on veille à la continuité des décisions en matière de sources et de dispositif d'indicateurs.

... tout en étant conscient des limites

Les artefacts sont nombreux en bibliométrie. Des risques majeurs sont associés :

- à la délimitation des périmètres et aux référentiels
- aux fluctuations statistiques notamment pour les indicateurs croisés (partenariats, etc.)
- à l'utilisation abusive des impacts et citations, notamment d'indicateurs non normalisés
- aux indicateurs de productivité très difficiles à mettre en oeuvre
- aux artefacts de représentation des méthodes statistiques.

ANNEXE I - SOURCES INTERNES ET EXTERNES

On peut attendre d'une institution de recherche qu'elle enregistre l'ensemble de sa production, quelle qu'en soit la nature (articles visibles internationalement, autre littérature scientifique, thèses, rapports, rapports d'expertise, etc.). La production est repérable dans deux types de sources, les sources internes et les sources externes. L'enregistrement de la production renvoie, pour les sources internes, aux systèmes propres d'information de l'institution, et pour les sources externes aux bases de données internationales sous leurs diverses formes.

1. Publications : sources internes

L'expérience montre qu'il est difficile de garantir l'exhaustivité d'un système interne de repérage des documents produits, y compris pour la littérature scientifique publiée dans les journaux internationaux. Une quasi-exhaustivité (pour la littérature majeure, publiée) ne peut être obtenue, en pratique, qu'en couplant le recueil de documents et les processus d'évaluation (des chercheurs, des unités etc). S'agissant de littérature "grise", les sources internes sont les seuls instruments utilisables. Si elles n'enregistrent pas ces documents, ils ne seront accessibles nulle part sous forme systématique (quelques traces volatiles sur Internet...).

La bibliométrie a des exigences complémentaires des "normes documentaires" applicables à une source, et ceci vaut pour les sources internes. De manière générale, les exigences sont convergentes : mieux une source satisfait les exigences des normes documentaires, plus elle est facilement utilisable en bibliométrie. La codification des contenus et des formats facilite grandement les analyses bibliométriques.

Toutefois, les exigences de l'analyse bibliométrique, qui véhicule en l'occurrence les besoins en termes d'étude stratégique, ne sont pas toujours satisfaites. Un problème bien connu est la mention de l'ensemble des adresses institutionnelles pour les acteurs présents dans une publication en co-signature - précision actuellement absente dans les sources internes INRA. Cet élément est pourtant absolument nécessaire pour les études de coopération scientifique.

Outre les caractéristiques bibliographiques habituelles qui assurent le lien, avec les bases externes spécialisées (essentiellement bases de publications et brevets), les éléments suivants devraient être disponibles pour des **analyses d'acteurs et de thématiques** :

Au minimum :

- toutes les adresses d'auteurs, au moins sous forme brute pour permettre des traitements ultérieurs, en indiquant si possible, la correspondance auteur/ adresse (sinon il faut prévoir une table chronologique d'affiliation des auteurs)
- identifiant unique des unités INRA participantes à la date de publication pour lien avec le/les référentiels "structures" (tables chronologiques sur les structures institutionnelles : unités, départements et centres de recherche)
- type de production, compatible avec le référentiel "évaluation " (ex grille CSS)
- titre et résumés en français et anglais pour permettre des traitements ultérieurs et la projection dans des référentiels variés

Informations complémentaires souhaitables

Elles peuvent être fournies lors de l'alimentation de la base, récupérées par liens avec des référentiels ou systèmes d'informations internes ou générées par des traitements plus ou moins automatisables. Dans ce dernier cas, il serait préférable de traiter l'ensemble des notices d'une année de publication en une seule passe pour des raisons d'homogénéité

- Décomposition et standardisation des lignes d'adresse pour faire apparaître les différents niveaux structurels et géographiques.

L'objectif est d'obtenir une information standardisée sur toutes les co-publications ainsi que sur les partenariats institutionnels (UMR ou autres). On conserve pour mémoire l'information d'origine.

Supposons qu'on récupère informations suivantes pour une notice :

Adresses brutes

A Inst Natl Agron Paris Grignon, Ctr Biotechnol Agroind, INRA, Chim Biol Lab, F-78850 Grignon//France
B INRA, Lab Biochim & Technol Prot, F-44316 NANTES 3//France
C Univ Grenoble 1, CNRS, Ctr Rech Macromol Vegetales /F-38041 Grenoble 9//France

Auteurs : AU1 (A), AU2 (B), AU3 (B), AU4 (C)

Exemple de structure de données pour les traitements bibliométriques

N° notice	N° auteur	Auteur	N° adresse	N° Orga-nisme	Organisme	Laboratoire	Ville	Code Postal	Pays	Codes INRA *	Ligne brute
1	1	AU1	1	1	INRA	UMR DE CHIMIE BIOLOGIQUE	THIVERVAL-GRIGNON	78850	FRA	TPVTG	A
1	1	AU1	1	2	INA PG	UMR DE CHIMIE BIOLOGIQUE	THIVERVAL-GRIGNON	78850	FRA	TPVTG	A
1	2	AU2	2	1	INRA	UNITE DE BIOCHIMIE ET TECHNOLOGIE DES PROTEINES	NANTES	44316	FRA	TPVNT	B
1	3	AU3	2	1	INRA	UNITE DE BIOCHIMIE ET TECHNOLOGIE DES PROTEINES	NANTES	44316	FRA	TPVNT	B
1	4	AU4	3	1	CNRS	CERMAV	GRENOBLE	38041	FRA	RORFRA	C
1	4	AU4	3	2	UNIV GRENOBLE 1	CERMAV	GRENOBLE	38041	FRA	EDUFRA	C
2	1	AU1	1								
2	1	AU1	2								

* ici AGEDOR ou PSEUDO-AGEDOR (type d'organisme+pays)

On peut ajouter: numéro codique INRA, centre INRA, département INRA, département géographique, type d'organisme,....

- Caractérisation "à gros grain" du contenu des documents pour les suivis thématiques chronologiques, les analyses rapides de profils et de spécialisation.

Pour le sous-ensemble des articles des journaux de la base ISI (WOS), on peut utiliser les spécialités scientifiques que leur affecte l'ISI. Les indicateurs "macro" de production scientifique élaborés pour la France par l'OST et les tableaux de bord par département et centre fournis par le LERECO-ESR. utilisent cette caractérisation, qui pourrait être intégrée dans un *référentiel "journaux"*.

Pour pouvoir caractériser l'ensemble de la production., un/ des codes de classement par document, à choisir dans un *référentiel externe multidisciplinaire reconnu et maintenu* (à définir, p.ex PASCAL / FRANCIS) pourraient faciliter les analyses rapides. Ces codes peuvent être fournis à l'alimentation ou affectés par traitement ultérieur si l'on dispose d'un outil de catégorisation des documents (voir p.ex les modules de STATANALYST de l'INIST).

Une troisième voie consisterait à exploiter un système d'indexation englobant tous les systèmes d'information interne, par exemple à des fins de navigation.

- **Représentation "fine" du contenu** des documents requise pour les structurations thématiques endogènes sur base lexicale ou citationniste (cf. ***Annexe II, référentiels thématiques)

Ce type de structuration ne s'impose pas toujours pour les tableaux de bord. Pour le positionnement stratégique, il s'appliquera au coup par coup à des corpus débordant largement les publications INRA et dépendra du contenu de ces corpus

Pour les seuls besoins bibliométriques internes, une procédure d'indexation automatique ou assistée assurerait un traitement homogène des documents tout en permettant de suivre l'évolution du langage. La combinaison de modules linguistiques et statistiques permet d'obvier, dans une certaine mesure, à une des principales limites des traitements automatiques du langage, la reconnaissance des variantes terminologiques chimiques, médicales, taxonomiques....

- **Eléments de visibilité**

Le référentiel "journaux" pourrait inclure des éléments robustes issus d'analyses internes à l'INRA (p.ex caractérisation de M.H. Magri) ou accessibles si un accord est passé avec des organismes extérieurs (p.ex à l'OST internationalisation, impacts divers).

Pour les données de visibilité par document, voir ci-dessous sur le WOS.

Fonctionnalités indispensables

Des modules d'import/export personnalisables pour la communication avec des SGBD et divers logiciels (bureautiques, bibliographiques, statistiques, data-mining...) sont indispensables. L'enrichissement par les utilisateurs pour leurs besoins propres de traitement (p.ex : axes stratégiques, champs thématiques, actions de recherche....) doit être prévu.

Pour les structures de données, il faut raisonner le niveau de redondance : il n'est pas souhaitable de dupliquer toutes les informations nécessaires dans la base "production" mais il faut aussi éviter de trop nombreuses opérations d'accès aux divers référentiels internes et externes.

Les modalités informatiques, l'organisation de l'alimentation de la base (qui peut/doit fournir quelle information, à quel stade de la chaîne etc ...), les niveaux d'affichage, d'accès etc... doivent faire l'objet d'une réflexion coordonnée des unités concernées à la DISI, notamment IST-PRODINRA et S2I-URAP.

2. Publications : Sources externes

21 types de bases

Les sources externes présentent souvent, paradoxalement, moins de "silences" (documents non repérés) incontrôlables que les sources internes. Plus exactement, les silences des bases de données sont en général systématiques et explicites, et liés aux choix de couverture de la base. Un bon exemple est la source ISI (Thomson Scientific). L'acquisition du Web of Knowledge/Web of Science permet une couverture quasi-systématique d'un large périmètre de journaux, sachant que ce périmètre est trop étroit sous certains angles (mauvaise couverture de certains domaines spécialisés) et trop large sous d'autres (présence de journaux orientés nationalement). PASCAL est la seule autre base multidisciplinaire à champ affiliation "bibliométrique" (toutes les adresses renseignées), sans décompte de citations mais avec indexation et plan de classement. Les choix de couverture de PASCAL ont malheureusement varié au fil du temps.

Pour des études de positionnement thématique fin, le recours à des bases thématiques spécialisées complémentaires peut être nécessaire, certaines accessibles en standard dans le dispositif INRA

(FSTA, Medline-Pubmed, E-Biomed en biomédecine), d'autres non (Inspec en physique, CAS-Chemical Abstracts en chimie-biochimie, BIOSIS en biologie, Compendex en sciences de l'ingénieur, etc.). Contrairement à l'ISI, certaines de ces bases comportent des descripteurs d'indexation sophistiqués (Mesh de Medline, index CAS, subject-headings CAB etc.), mais ont certains inconvénients (pas toujours de saisie intégrale des journaux, pas de saisie complète des affiliations, pas de citations).

Les relations de citation sont accessibles en navigation sur certaines bases spécialisées (p.ex. CiteSeer en informatique). Les champs adresses se limitent souvent à la première affiliation et ne sont pas nécessairement de grande qualité. Ces bases se prêtent donc moins à l'analyse des réseaux.

Les modes de mise à jour et d'accès importent pour l'utilisation bibliométrique. Les mises à jour très réactives implémentées par exemple sur la version WoS des bases ISI, et l'ergonomie du produit adaptée au Web, rendent cette version très utilisable en navigation/ veille, mais très difficilement utilisable pour les analyses de positionnement sans un déchargement massif des données.

22 utilisation du WoS en bibliométrie

L'utilisation bibliométrique des bases dépend, outre leur contenu, de leurs interfaces. Ainsi le WoS/WoK dans sa version standard (internet ou intranet) est orienté vers la navigation rapide ce qui en fait un outil remarquable pour la finalité "veille". Pour les finalités "tableau de bord" ou "positionnement stratégique" c'est un instrument assez peu adapté, instable (mises à jour permanentes) et ne permettant pas d'agrégations commodées. Il faut donc dans ce cas procéder à des déchargements de la source, mais les conditions de déchargement sont elles-mêmes difficiles (nécessité de diviser les requêtes pour éviter la saturation en version Internet).

Une solution alternative est le partenariat avec des détenteurs des versions "à plat" orientées bibliométrie et qui n'ont pas ces contraintes (telle que la version négociée à l'OST).

L'ISI propose depuis quelques années des produits d'indicateurs, sous forme désagrégée (liste de publications sélectionnées sur un critère donné, p.ex. un pays ou une institution) ou agrégée (p.ex. "National science indicators" au niveau pays). Une forme plus souple, interactive, vient d'apparaître comme extension du WoK, les "Essential science indicators" qui fournissent rapidement les cumuls par acteur sur un domaine.

Les principales limitations de ces produits sont :

- l'absence de flexibilité des indicateurs calculables et des filtres mis en oeuvre
- l'absence d'unification des données, une limitation très sévère (voir annexe)

23 PI et brevets

Le brevet est plus codifié que la publication scientifique, mais le principe de "révélation minimale" d'information rend certaines descriptions (titres) peu utilisables. Les "références citées" (à d'autres brevets, aux publications, etc.) sont exploitables et ont renouvelé l'économie de la S&T, mais leur interprétation varie selon les systèmes et reste délicate.

Chaque système de brevets a sa base de données publique. A l'échelle internationale, EPAT et USPAT (et à un moindre degré la base japonaise) sont les plus utilisées. Notons que l'OST dispose d'une version quasi-complète de la base EPAT. La base multi-système Derwent/WPI (Thomson), avec familles, titres recomposés, nomenclature détaillée, est coûteuse mais performante pour les analyses bibliométriques. Les bases nationales (ex. INPI) sont mobilisables au cas particulier.

La base COV se prête difficilement, en l'état, à des études bibliométriques sophistiquées.

24 autres sources.

Diverses sources hors publication sont exploitables pour l'analyse stratégique ou le suivi des politiques institutionnelles, par des méthodes bibliométriques. Exemple :

- les contrats
- les missions
- les CR d'évaluation
- la politique de communication (sites web)
- etc...

ANNEXE II METHODES ET INDICATEURS

Les indicateurs bibliométriques courants sont analogues à ceux rencontrés en économie industrielle et internationale, sur la base d'analogies entre production scientifique et production de biens et services. L'information scientifique peut être considérée comme un bien public aux propriétés particulières.

Les spécificités des indicateurs bibliométriques et de leur interprétation tiennent essentiellement :

- sur le plan statistique, au caractère très inégal des distributions rencontrées, qu'il s'agisse de productivité individuelle, de concentration des sources documentaires, de concentration des descripteurs p.ex. linguistiques. Cette concentration est un principe efficace de sélection - il suffit d'une fraction de l'information disponible pour décrire convenablement un univers - mais compromet la robustesse des approches statistiques standards, car un très petit nombre d'individus ou d'items peuvent altérer profondément une répartition. Ce type de distribution se rencontre également en économie.

- à l'existence d'un réseau particulier, celui des "citations" entre articles construit à partir des références bibliographiques. Ces citations permettent de mesurer la visibilité internationale des travaux scientifiques, et, concurremment avec les marqueurs linguistiques, procurent des moyens d'investigation puissants des réseaux S&T. Les citations existent aussi dans les brevets, avec des significations différentes. La citation est un phénomène très complexe et les indicateurs fondés sur les citations exigent de grandes précautions dans l'interprétation.

Les référentiels thématiques sont d'une importance essentielle en bibliométrie. On peut distinguer "macro" et "micro" bibliométrie selon la nature de ces référentiels : on parle souvent de micro-bibliométrie lorsque l'unité de délimitation des périmètres est l'article, de macro-bibliométrie lorsque l'unité est le journal entier.

On distingue :

- les référentiels externes, nomenclatures ayant une certaine permanence et/ou une certaine diffusion. Comme il n'existe pas de classification des sciences universellement admise, on recourt soit aux définitions "macro" de l'ISI (chaque spécialité définie par une liste de journaux), soit aux définitions (macro ou micro) proposées par les plans de classement d'autres bases de données (multidisciplinaire comme Pascal, ou thématiques). Pour les brevets, la situation est différente, avec deux nomenclatures officielles de base (Classification Internationale des Brevets, à base européenne, Classification USPTO).

- les référentiels ad hoc, soit macro soit micro, construits par exemple par méthode bibliométrique (mots associés, réseaux de citations, co-classements, etc.). Certaines nomenclatures sont elles-mêmes partiellement fondées sur des analyses bibliométriques.

- certains référentiels de classement ne sont pas à proprement parler thématiques, mais stratégiques ou politiques : axes prioritaires, finalités sociales, etc. Ces référentiels externes peuvent être acclimatés en tant que de besoin.

1. indicateurs bibliométriques classiques

a) option : type de comptage

Le choix du dénombrement pour les articles à adresses multiples détermine les mesures bibliométriques. Le type de comptage standard en micro-analyse est le compte de présence (1 si

l'institution ou l'acteur est présent dans la publication, 0 sinon). Son principal inconvénient est l'instabilité dans les agrégations, du fait des comptes multiples (co-publications). Les sommes sont en général supérieures à 100% (la somme des publications par département de recherche est supérieure au nombre de publications de l'institution ; la somme des publications par région est supérieure au nombre de publications françaises, etc.). L'avantage est la facilité d'interprétation, notamment pour les co-publications, et la stabilité par rapport à l'information disponible sur les acteurs d'une publication. Le compte fractionnaire (adapté aux approches macro, et consolidable à toutes les échelles) peut être également pratiqué dès lors que le repérage s'effectue par adresse distincte.

b) production et profil de spécialisation

concentration et spectre de publication dans les journaux du périmètre
positionnement de l'acteur en nombre d'articles, "parts de marché" française, européenne, mondiale
profil de spécialisation sur des référentiels thématiques ou stratégiques (grilles disciplinaires, axes de prioritaires, etc.)

c) visibilité sur plusieurs critères

Les indicateurs de visibilité sont riches mais d'interprétation très périlleuse. Leur emploi nécessite un certain professionnalisme. Le premier danger est de négliger les particularités dans les comportements de citation, très différents selon les champs. Un seul indicateur non normalisé, comme l'"impact factor", est inutilisable. Une batterie d'indicateurs variant les points de vue et les niveaux de normalisation est nécessaire. Quelques indicateurs :

"part de marché des citations"
impact brut, impact relatif, (en valeurs directes et normalisées par champ)
impact espéré et ratio de citation relative (stratégie de publication) ; répartition des publications selon l'internationalisation des journaux.
distribution des publications par tranche de visibilité et présence dans les tranches d'"excellence" (il existe d'autres indicateurs d'"excellence", que nous ne détaillons pas ici).

d) coopérations internationales/ inter-régionales

intensité et structure globales de collaboration ; répartition France/ Europe/ Autres
profil par pays, par région (si pertinent), par institution française (pour les données partageables de la coopérative⁷)
collaborations intra-institution sur les entités identifiées (entre départements de recherche, entre centres, etc.)
à terme (sur Web of Science), profil des partenariats par acteur (institutions, firmes, etc.), avec standardisation des adresses sous la responsabilité de l'institution.
variété des points de vue : transactions brutes, transactions corrigées de la taille des acteurs (indices d'affinité mutuelle), mesures intermédiaires de proximité.

e) productivité

Les entrées (financières, ressources humaines...) sont en principe inaccessibles dans les bases de données de "sorties" (publications). La bibliométrie se borne le plus souvent à explorer les "performances" des chercheurs actifs, mais à grande échelle les problèmes d'homonymie ou de synonymie au sens large rendent l'exercice très difficile. Dans une approche plus générale de la productivité (en termes de ressources humaines de divers types, en termes de ressources financières), les données d'input des acteurs sont indispensables, avec de sérieux écueils.

⁷ supposent un accord particulier avec les institutions du groupe pour un partage des fichiers de repérages d'adresses.

- absence de comparabilité à l'échelle individuelle (budget temps, cursus individuels, etc.) et difficulté d'établir des références moyennes
- types de ressource humaine comptabilisée (affectations des personnels techniques, administratifs...).

La productivité en termes de ressources financières posent des problèmes analogues (pratiques comptables, problèmes d'affectation régionales/ nationales etc.). La comparaison des "productivités" entre disciplines risquent aussi d'être dépourvues de sens (coût des équipements, nature des publications, etc.).

f) liens science-technique

Pour mémoire, la bibliométrie est un outil privilégié, à côté des enquêtes directes, pour mesurer les liens science-technologie (et université-industrie) à diverses échelles : macro (liens entre nomenclatures), méso (positionnement de firmes ou de régions), micro (liens entre publications et brevets). L'ensemble des réseaux habituels (lexicologie, auteurs - notamment ici le cas de co-activité, citations) est mobilisable pour l'analyse.

g) combinaison d'indicateurs

Un indicateur pris isolément peut être dépourvu d'interprétation claire et seule une variété de points de vue permet de décrire l'activité complexe des acteurs, même si cette fraction d'activité que constitue les "publications académiques"...

2. indicateurs de réseau

Les principaux réseaux analysés en bibliométrie sont les réseaux d'acteurs (auteurs, institutions, pays...), les réseaux de citation, les réseaux lexicologiques au sens large. Des méthodes de graphe (réactivées dans ces applications par la théorie des réseaux sociaux) et d'analyse de données permettent d'identifier des zones de concentration (thèmes, groupes d'acteurs, etc.) et des relations entre individus ou groupes. Par exemple :

- densité et centralité locales, popularisées à la fin des années 80 par le "diagramme stratégique" de Leximappe (attention toutefois à l'interprétation de ces mesures sur des thématiques de recherche : bien vérifier la stabilité des mesures dans les changements d'échelle)
- les synonymies/ polysémies-homonymies peuvent s'approcher en termes de structures de lien (équivalences dans le graphe pour les synonymies au sens large, passerelles entre contextes pour les polysémies). L'analyse de réseau rejoint ici l'arsenal des techniques d'analyse du langage.
- chemins indirects (propriété de "small world" avec groupes denses mais connexions plus lointaines permettant des propagations générales), etc. Ceci peut être mis à profit pour l'analyse des fronts de recherche, aussi bien que des acteurs individuels.

A l'échelle globale, ce sont les mêmes matrices de liens de co-occurrence, ou tableaux de profils, qui sont mises en oeuvre pour la structuration globale du domaine, après sélection d'une relation structurante (voir ci-dessous).

3. mise en oeuvre d'une analyse bibliométrique de positionnement

Une analyse typique de positionnement met en jeu les commanditaires, les experts du champ, les bibliomètres et les spécialistes des sources. Le commanditaire est central dans la phase de définition des objectifs et de discussion/exploitation des résultats, l'interaction entre experts et bibliomètres est nécessaire dans la plupart des phases techniques d'élaboration.

a) définition des objectifs

Par hypothèse ici, l'analyse est centrée thématiquement mais naturellement les besoins du commanditaire peuvent être différents (p.ex. décrire l'activités de partenaires-concurrents).

b) choix des sources et délimitation du champ.

Le choix des sources est évidemment capital.. Pour la production scientifique on peut trouver différents cas de figure :

- domaine bien couvert par une seule source documentaire avec une bonne indexation/terminologie familière aux demandeurs : structuration lexicale sur mots-clefs possible, récupération souhaitable des adresses des auteurs dans d'autres bases (PASCAL, ISI). Idem pour les citations si on veut combiner méthodes lexicales et citationnistes.

- domaine multidisciplinaire et couvert par les périodiques ISI : structuration citationniste possible, combinée au moins avec une verbalisation par traitement des titres, ou avec une analyse lexicale sur vocabulaire naturel ; dans certains cas les indexations PASCAL pourront être mobilisées.

- domaine multidisciplinaire avec des aspects mal couverts par les bases ISI ou PASCAL/ FRANCIS (p.ex SAD ...) le traitement du langage naturel s'impose.

Cette dernière configuration se retrouvera pour tout corpus hétérogène, en particulier si l'on veut couvrir différentes dimensions des activités de l'Institut.

La phase de détection des articles (ou brevets) pertinents est tout à fait cruciale et exige la combinaison de compétences documentaires et bibliométriques. Le but est d'obtenir un bon "rappel" (la majorité des articles pertinents seront repérés) et une bonne "précision" (peu d'articles non pertinents seront renvoyés).

La phase de délimitation conditionne la pertinence de l'analyse ultérieure.

- méthode documentaire traditionnelle : le contour d'un thème s'approche souvent selon les étapes suivantes : identification des journaux spécialisés et extraction extensive de leur contenu, identification d'articles de synthèse, puis définition de requêtes lexicales, d'après les bonnes pratiques de la théorie documentaire, en utilisant par exemple comme germe le vocabulaire des journaux spécialisés, des articles de synthèse, et des experts du domaine associés à l'étude. La forme des requêtes dépend des bases interrogées et de ses interfaces (requêtes booléennes, moteur de recherche sophistiqué, etc.).

- extension aux marqueurs de citation. Les citations peuvent s'avérer plus puissantes que les marqueurs lexicaux pour délimiter des domaines transversaux ou émergents, au vocabulaire non stabilisé.

- méthodes statistiques/ bibliométriques : à un premier niveau, la détection de réseaux de proximité, par exemple de co-occurrences (mots associés, couplages, co-citations), permet d'affiner les requêtes ci-dessus en réduisant notamment les problèmes de synonymie par exemple. Des techniques d'analyse de pertinence peuvent également être mises à profit pour affiner les périmètres. A un second niveau, ces méthodes peuvent être utilisées pour une première cartographie globale du domaine, p.ex. par voie factorielle ou de classification, afin notamment d'éliminer des groupes d'articles non pertinents, remontés par effet de polysémie/homonymie. Un traitement du langage naturel (voir ci-dessous) peut être nécessaire en amont.

- ces approches peuvent être chaînées en séquences itératives, enchaînant par exemple une délimitation par excès sur une formule lexicale large et un affinement par repérages des articles structurants fortement cités, puis à nouveau une sélection lexicale. Le chaînage optimal est largement fonction des caractéristiques des champs.

- l'intervention des experts du champ est capitale, aussi bien dans la détermination du contenu des requêtes, leur affinement ultérieur, que dans la validation finale du champ. Chaque étape peut faire l'objet de supervisions par les experts.

c) structuration du champ

Le problème est analogue à la délimitation de micro-champs (thèmes), mais il serait très coûteux d'identifier individuellement chaque thème par les méthodes itératives décrites ci-dessus. On fait donc confiance aux méthodes de la statistique, de l'infométrie, de la fouille de données, de la bibliométrie pour structurer les champs. Mais en fonction de la demande, cette structuration peut reprendre tout ou partie des itérations précédentes, et faire l'objet de supervisions par les experts ou le commanditaire. Une validation finale est toujours nécessaire, en raison des artefacts des méthodes bibliométriques (couverture du champ, artefacts statistiques).

Relations structurantes :

Limitons-nous aux plus classiques, les méthodes lexicologiques et citationnistes (pour les brevets en particulier, on recourt également au co-classement). D'autres liens sont théoriquement utilisables (co-publication) mais plus rarement utilisés pour cette application particulière.

- **méthodes lexicologiques.** Elles sont universelles, mais dépendantes des langues. Leur application est sensiblement différente selon que l'on analyse un langage contrôlé de bonne qualité (Mesh, index de CAS) ou le langage naturel des titres, résumés ou texte intégral, où se posent de sérieux problèmes d'unification : flexions, synonymies, homonymies, abondance d'abréviations et acronymes, etc.

Les méthodes de traitement du langage naturel qui ont connu un développement considérable dans la décennie peuvent être mobilisées pour l'extraction terminologique, avec des fonctionnalités variées : lemmatisation, détection des multi-termes, unification des concepts, vocabulaires relatifs (différences de distribution de vocabulaire dans un domaine spécialisé et dans le langage courant) etc. Le traitement de certains textes scientifiques (exemple les résumés d'article en chimie) est un défi à ces techniques, fondées sur une panoplie d'approches : analyse morpho-syntaxique, réseaux sémantiques, méthodes de graphe, méthodes statistiques.

L'analyse lexicale n'est pas dynamique par construction, comme peut l'être l'analyse des citations reliant deux articles de dates différentes. L'analyse dynamique est possible par le suivi longitudinal des fréquences de termes ou d'associations, l'étude structurelle prédictive, la statique comparative sur les clusters. L'étude des métonymies et des évolutions des familles de termes caractérisant les émergences thématiques est prometteuse mais difficile. Si l'on utilise un vocabulaire contrôlé, il est indispensable de calibrer les vari

Deux voies typiques pour l'approche lexicale sont l'analyse des co-occurrences ("mots associés" sur divers indices de similarité, suivie typiquement d'une classification automatique ou d'une MDS) et l'analyse factorielle (AFC, Latent Semantic Indexing) ou apparentée (K-means axiales). D'autres voies sont naturellement utilisables (classifications neuronales, etc.).

Les représentations cartographiques sont propres à chaque voie d'analyse :

- représentations discontinues (graphes à seuil, arborescences, etc.) pour les relations de similarité aboutissant à l'identification de "thèmes de recherches". Idem pour les diagrammes à cases fournies par les classifications neuronales.
- représentations continues sur un ou plusieurs plans pour les analyses factorielles. Des représentations plus originales se développent à la faveur des techniques de cartographie du Web.

Si les techniques factorielles sont riches en interprétation par la mise en évidence de dimensions latentes, il est utile d'aboutir *in fine* à une représentation discontinue en "thèmes de recherche" afin de les caractériser par une batterie d'indicateurs bibliométriques.

- **méthodes de citation.** Leur champ d'application est limité en pratique aux domaines couverts par les bases avec références citées (ISI, CiteSeer...; bases de brevets). Elles sont, au premier ordre, indépendantes des langues (mais des biais existent naturellement), et ne posent évidemment pas les problèmes d'extraction terminologique propres au langage naturel. La relation de citation est diachronique et se prête donc à des analyses dynamiques d'application des connaissances.

Les caractéristiques infométriques des citations en font un outil très riche de représentation des univers scientifiques, en tenant compte de leurs caractéristiques particulières bien analysées par les sociologues et les bibliomètres. Un avantage des citations est leur aptitude à décrire la complexité : les distributions statistiques sont très concentrées, mais à moindre degré que le vocabulaire naturel, elles sont donc l'équivalent d'un vocabulaire "extrêmement riche".

Les techniques d'analyse et les représentations cartographiques sont sensiblement les mêmes que pour les études lexicales. Les applications les plus classiques suivent la voie des co-occurrences, co-citation et couplage bibliographique, mais rien n'empêche l'emploi de techniques continues. Lorsque la structuration par co-citation conduit à des clusters, on parle souvent de "fronts de recherche" (qui ont un côté cité et un côté citant) par référence aux premières études de l'ISI.

d) indicateurs thématiques/ globaux

Dès lors que le niveau et le périmètre des "thèmes" ou "fronts" de recherche sont validés, il est utile de cartographier l'agencement des thèmes au sein du domaine étudié sur la relation structurante.

L'analyse micro-bibliométrique caractérisera ensuite les acteurs thème par thème sur une série d'indicateurs, notamment ceux répertoriés plus haut. Parallèlement, les profils et spécialisations pourront être établis par rapport aux thèmes identifiés par la structuration.

Dans la mesure où l'identification des thèmes a été obtenue par une relation structurante particulière, il peut être utile de compléter (ou comparer) ces liens et cette représentation par les relations secondaires, par exemple des liens lexicologiques si on est parti d'une étude citationniste.

La comparaison entre clusters et cartes issues de méthodes différentes est d'ailleurs particulièrement intéressante. La complémentarité entre méthodes lexicales et de citation n'est pas pleinement exploitée dans la plupart des études.

e) discussion-exploitation

C'est le moment pour les bibliomètres d'attirer l'attention sur les limites de l'exercice, rappelées dans le rapport. Comme dans d'autres techniques d'aide à la décision, on a coutume de dire que 80% des résultats d'analyse confortent le commanditaire dans sa connaissance, que 10% sont des artefacts, et que 10% apportent des éléments intéressants de réflexion. C'est une ambition déjà honorable...

ANNEXE III LA BASE LOGICIELLE

Pour les finalités ci-dessus, les techniques bibliométriques mobilisent un ensemble très large de techniques et de logiciels pour l'extraction et le stockage de données, l'analyse linguistique, les traitements et statistiques et analyses de données, les représentations graphiques et la mise à disposition des résultats.

L'intégration de ces fonctionnalités dans une architecture modulaire permet la combinaison flexible de diverses formes d'analyse bibliométrique, associant stockage rigoureux de l'information (base de données relationnelle ou objet), facilités de navigation et mise en œuvre de méthodes statistiques robustes. Dans l'état actuel de l'offre, ce modèle peut-être approché par la combinaison de logiciels commerciaux interopérables, complétée par des développements spécifiques ou des logiciels libres.

Un modèle possible intègre quatre types d'éléments:

1- une base de données, typiquement relationnelle, au centre du dispositif, pour la navigation à la fois dans les données et dans les résultats. Cette base peut s'adosser en amont les systèmes d'information interne

2- des modules d'extraction et de reformatage de données, modulables selon les types de sources et les types de traitement envisagés (sur texte entier, champs textes ou autres de fichiers structurés ...). L'interaction avec les modules d'analyse est souhaitable pour la reformulation des requêtes, enrichissement de listes d'autorité pour les reformatages etc....

3- des modules d'analyse linguistique et statistique intervenant à divers stades des traitements. Ces modules chargent typiquement leurs données sur la base, et y réinjectent les résultats. Par exemple, la réinjection de clusters dans la base permet, sur ces clusters, une variété d'analyses (co-publications, analyse thématique, etc.).

4- Interface Homme-Machine: modules de navigation sur la base, à fonctionnalités web. Ces modules sont d'une part des modules de navigation classiques (dans les tables de données/ résultats) et des modèles graphiques de préférence interactifs (cartes "clicquables"). La séparation des fonctions 2 et 3 dépend des solutions logicielles mises en œuvre.

Cette architecture est représentée dans la figure ci-dessous.

L'analyse de l'offre logicielle doit prendre en compte les fonctionnalités de traitement offertes dans la gamme utile pour l'analyse bibliométrique ainsi que les éléments classiques rappelés pour mémoire :

- type de technologie, architecture, plate-forme ;
- modularité et capacité d'échanges
- flexibilité, adaptabilité/ personnalisation, évolutivité
- transparence (boîtes noires...) et robustesse des traitements
- capacité de traitement
- niveau d'utilisateur requis
- ergonomie, Interfaces Hommes Machines pour le différentes parties prenantes (ingénieurs documentaires, experts du domaine, bibliomètres, demandeurs divers ...)
- statut propriétaire/ libre, universitaire/ commercial, type de commercialisation (vente/ location/ ou support d'études non cédés...)

Une analyse comparative de l'offre nécessiterait des essais comparatifs sur jeux de données et protocoles définis. La mise au point et le test de protocoles de bibliométrie exploratoire n'est réalisable à court terme que pour les instruments disponibles à l'INRA dont SAS et SPHINX.

Le premier objectif du groupe a été de recueillir des avis d'utilisateurs. Quelques logiciels ont fait l'objet de présentations au groupe (voir ***Annexe IV, documents disponibles), d'autres ont été utilisés à l'INRA et présentés dans d'autres instances, notamment des applications de Tétralogie et Leximine.

Dans la gamme légère, les fonctionnalités linguistiques et statistiques de Sphinx en font un outil recommandable. Ses autres fonctionnalités sont également intéressantes (traitement d'enquête, reporting, etc.). La prise en main est relativement rapide. Un outil complémentaire d'ailleurs interfacé avec le précédent, est Alceste, qui structure un corpus par classification descendante. Dans un registre proche, Diatopie utilise les k-means axiales.

Tétralogie est un logiciel de statistique générale orientée bibliométrie, plutôt axé sur des sources hétérogènes, avec reformatage assez rustique, mais des modules complémentaires peuvent sans doute améliorer la gestion des données. Les représentations graphiques à base d'analyse factorielle et de classification sont agréables, ainsi que le module de cartes à seuil. Le logiciel est généralement considéré comme délicat pour des utilisateurs occasionnels, et la société productrice se positionne le plus souvent en prestataire du service associé. Le logiciel fait partie de la moyenne gamme, sans le niveau de complexité et de contrôle offerts par les poids lourds du data-mining, en particulier SAS et SPSS.

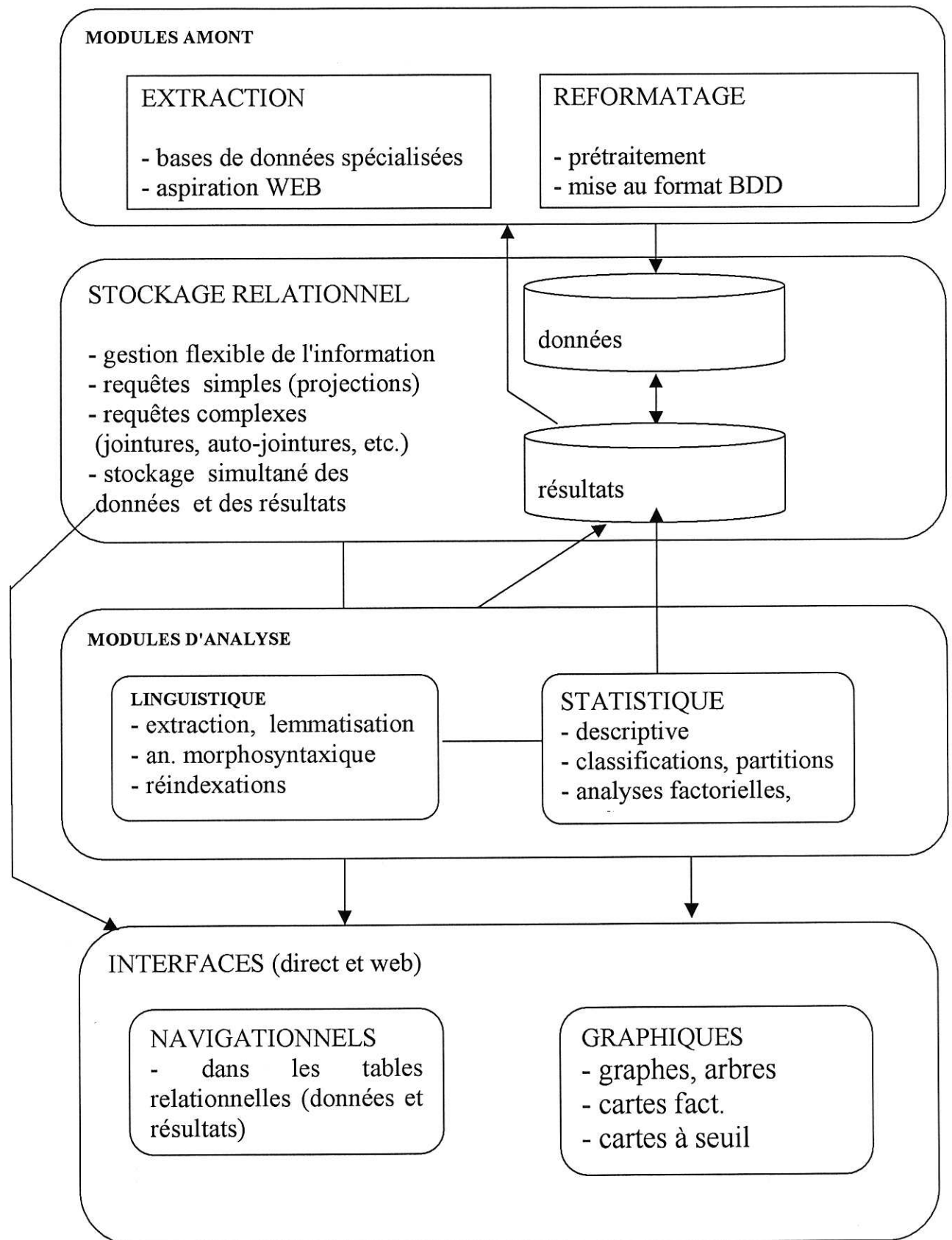
Comme le précédent, Statanalyst de l'INIST-CNRS est souvent présenté en prestation de service, il intègre plusieurs modules de traitement linguistique associés à des techniques de structuration de type association et k-means axiales. Il est adossé aux bases de l'INIST et des tests de cet outil seraient recommandables malgré certaines limitations.

Leximine est un des modules de l'offre SPSS « Clementine ». C'est un outil de navigation et de veille utilisant les liens de co-occurrences homogènes (auteurs-auteurs, mots-mots...) ou hétérogènes (auteurs-mots, etc.). Il permet de naviguer efficacement dans un corpus de très grande dimension (typiquement non structurés, mais l'adaptation était à l'étude). Les fonctionnalités de clustering sont très limitées (liens ternaires) et renvoient plutôt aux modules de classification de Clementine. Des séries fréquentielles chronologiques sont implémentées, mais ce n'est pas vraiment un outil de statistique bibliométrique. D'après les utilisateurs, le logiciel demande une certaine expertise pour être pleinement exploité.

Pertimm représente une autre famille d'outils, très orienté veille. C'est un moteur d'exploration de sites et données Web, avec des fonctions d'interrogation particulièrement sophistiquées, mais qu'il est probablement difficile d'exploiter pleinement sans la possibilité de développer des modules informatiques spécifiques.

Il existe une panoplie de logiciels statistiques généraux (SPAD avec fonctions de text-mining, STAT-A, etc.). Parmi ceux-ci, SAS et SPSS se distinguent par la richesse des fonctionnalités, dont certaines accroissent considérablement le coût. Les modules de base de SAS sont intégrés à la boîte à outils INRA sur Unix, et disponibles en version micro sur la base de l'« offre académique » de SAS. La version de base intègre le langage SQL avec la puissance d'un SGBDR (mais sans les fonctions de sécurité et de temps réel). De nouveaux modules dont Entreprise Guide (sur micro) ont grandement amélioré l'interface et permettent l'accès à SAS d'utilisateurs moins avertis. SAS offre de très nombreux modules (économétrie, etc.). Un nouveau module data-mining est disponible (Entreprise Miner, intégrant optionnellement l'outil linguistique Word Miner), à un coût très élevé.

SPSS a une offre du même type, avec la suite de data-mining Clementine, et l'outil de navigation dans les réseaux Leximine/Lexiquet mentionné plus haut.



ANNEXE IV DOCUMENTS DISPONIBLES

Documents de formation

Session mars-avril 2003

- *fiches techniques bibliométrie : liste indicative*

en première version	ébauchées	à venir
<i>Mode de comptage des publications</i> <i>Mode de comptage des co-publications</i> <i>Indicateurs de citation</i> <i>Indicateurs d'activité/spécialisation</i> <i>Indicateurs/ Volume de publication</i> <i>Indicateurs de copublication</i> <i>Impact et Impact Factor</i>	<i>Impacts relatifs/ normalisés I</i> <i>Indicateurs de journaux</i> <i>Mesures de concentration</i> <i>Mesures de dispersion</i> <i>Position dans le réseau</i> <i>Relation de citation</i>	<i>Généralités bibliométrie</i> <i>Lois bibliométriques</i> <i>Indicateurs de croissance</i> <i>Mesures d'excellence</i> <i>Structuration des univers scientifiques par les citations</i> <i>Structuration des univers scientifiques par les lexiques</i> <i>Méthodes de mots associés</i> <i>Impacts relatifs/ normalisés II</i> <i>Référentiels scientifiques</i> <i>Bibliométrie de brevets I</i> <i>Bibliométrie de brevets II</i> <i>Sources</i>

Fiches Mode de comptage des publications et Impact et Impact Factor.

MZ-EBZEX-fichV1.doc

- *jeu d'essai extrait du WOS*

jeu d'essai "Agricultural Economics", exemple de séquence de traitements sur logiciel SAS :

EBZ-MZ-jeu-essai.doc

Session décembre 2003

- *Exposé introductif bibliométrie : MZ-EBZ-ExpoBiblio.ppt*

- *Exposé introductif Sphinx pour la Bibliométrie LEISER-ExpoSphinx.ppt*

Présentations de logiciels

par des utilisateurs

ROUSSEAUHANS-compar.ppt (DISI-2002)

De l'extraction de l'information à la cartographie, présentation comparative de logiciels d'extraction, traitement et représentation d'information textuelle Excell, Simbad (Cea), Coelis, Tetralogie, Neurodoc (intégré à Statanalyst), Diatopie, Sampler (pour mémoire), Leximine, Kartoo.

BEAUDOIN-Leximine.ppt (DISI-2002)

Utilisation de Leximine pour les analyses bibliométriques

-pour les experts du domaine, Navigation, Identification de concepts, Veille

- pour les experts TI : Synthèses à l'aide des index, Prétraitement pour les analyses statistiques

par des producteurs

Diatopie.ppt (ADEST-2002)

logiciel développé sur la méthode des K means axiales d'Alain Lelu (associé INRA-MIA)

Statanalyst.ppt (IST-2003)

Suite modulaire de traitement de l'information développée par l'INIST-CNRS

Grille indicative d'analyse des projets

A partir des retours d'expériences plus ou moins formalisés et des discussions du groupe, une grille utilisable comme aide pour caractériser les expériences déjà conduites ou préciser les contours d'une nouvelle demande est proposée :

COLL-Grille-Projet.doc