UNIVERSITE DE DROIT, D'ECONOMIE, ET DES SCIENCES
D'AIX MARSEILLE
FACULTE DES SCIENCES ET TECHNIQUES DE SAINT JEROME

N° attribué par la bibliothèque
/_/_/_/_/_/_/_/_/_/_/

L'HYPERTEXTE COMME MODE D'EXPLOITATION DES RESULTATS D'OUTILS ET METHODES D'ANALYSE DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE

THESE

pour obtenir le grade de **Docteur en Sciences**
de l'Université de Droit, d'Economie et des Sciences d'Aix-Marseille
Discipline : Sciences de l'information et de la Communication
présentée et soutenue publiquement par

Luc GRIVEL

le 10 janvier 2000

JURY

M. Luc Quoniam, Professeur à l'IUT Service et Communication à St Raphael, Directeur de thèse

M. Jacky Kister, Directeur de Recherche au CNRS, Co-directeur de thèse

M. Jean-Francois Marcotorchino, Directeur du Centre Européen de Mathématiques Appliquées (CEMAP) d'IBM et Professeur associé à l'Université de Marne la Vallée

M. Thierry Lafouge, Maître de Conférence à l'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), habilité à diriger des Recherches

M. Xavier Polanco, Responsable de l'Unité Recherche et Innovation de l'Institut de l'Information Scientifique (INIST), CNRS

Remerciements

Je tiens à remercier toutes les personnes qui, par leur aide ou leurs encouragements, m'ont permis de réaliser cette thèse :

Alain Chanudet, directeur de l'Institut d'Information Scientifique et Technique (INIST-CNRS), qui m'a permis d'effectuer cette thèse dans le contexte de l'INIST, pour son soutien,

Henri Dou, responsable du Centre de recherche rétrospective (CRRM) de Université d'Aix Marseille III, pour m'avoir accueilli dans son laboratoire,

Luc Quoniam, professeur à l'IUT Service et Communication à St Raphael, qui m'a incité à effectuer cette thèse et m'a permis de la réaliser, pour son encadrement efficace et bienveillant,

Jacky Kister, Directeur de Recherche au CNRS, qui a co-dirigé cette thèse, pour son intérêt pour mes travaux, son soutien et ses encouragements,

Jean-Francois Marcotorchino, Directeur du Centre Européen de Mathématiques Appliquées (CEMAP d'IBM) et Professeur Associé, et Thierry Lafouge, Maître de Conférence à l'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), habilité à diriger des Recherches, pour avoir accepté la charge d'évaluer ces travaux, ainsi que pour leurs remarques,

Xavier Polanco, responsable de l'Unité Recherche et Innovation à l'INIST, avec qui je collabore depuis 10 ans, pour sa confiance, son soutien scientifique et moral,

Charles Huot, Responsable du segment 'fouille de données textuelles' à IBM, pour ses conseils et remarques,

tous les membres de l'URI, et notamment Dominique Besagni, Claire Francois, Jean Royauté, sans lesquels la majeure partie de cette recherche n'aurait pu se faire,

Jacques Ducloy, responsable de la première entité de recherche à l'INIST, qui m'a communiqué son expérience de la gestion de projets, et qui m'a fait confiance dans la conduite du projet SDOC,

tous les stagiaires pour leurs développements informatiques, et notamment trois élèves-ingénieurs qui ont participé en 1995 pendant 6 mois au projet HENOCH (Charles Broussaudier, Bruno Levy, André Kaplan), dans le cadre d'un stage de l'école supérieur en informatique et automatisme de Lorraine, (ESIAL).

Mes parents, Catherine, ...

Table des matières

Préambule

Cette thèse s'est déroulée dans le cadre d'une activité de recherche et développement que j'effectue depuis 10 ans à l'Institut d'Information Scientifique et Technique (INIST[1]), premier centre intégré d'information scientifique et technique en Europe dont la mission est, au sein du Centre National de la Recherche Scientifique (CNRS), de collecter, traiter et diffuser les résultats de la recherche scientifique et technique. Les évolutions du service chargé des activités de recherche et développement auquel j'appartiens expliquent certaines de mes orientations méthodologiques et informatiques. Ces changements, qui ont abouti à la création en 1998 de l'Unité Recherche et Innovation (URI[2]), m'ont permis de collaborer avec de nombreuses personnes. J'ai ainsi pu bénéficier de l'expérience acquise par le SERPIA[3], dirigé par William Turner au sein du CDST. Dans le cadre d'un projet européen (KWICK Esprit II project n°2466) initié par William Turner, j'ai eu la responsabilité de développer un outil nommé SDOC, basé sur la méthode des mots associés[4], fruit d'une collaboration entre le Centre de Sociologie de l'Innovation de l'Ecole des Mines de Paris et le CDST. Avec Jacques Ducloy, responsable du DRPN[5] de 1991 à 1993, j'ai participé au développement d'une boîte à outils pour le traitement de l'Information Scientifique et Technique. Avec X. Polanco, responsable de l'URI, je travaille depuis 1993 à la définition méthodologique et opérationnelle de l'analyse de l'information au sein d'une équipe de cinq ingénieurs double compétence (à la fois informatique et scientifique), spécialisés dans les sciences et technologies de l'information.

Les travaux qui sont présentés dans cette thèse se situent dans le cadre du développement d'une plate-forme logicielle[6] dédiée à l'analyse de l'IST. Ce développement, qui a débuté en 1993, se poursuit actuellement au sein de l'URI sous la forme d'une station de travail intégrée nommée STANALYST® (marque déposée). Ces travaux ont donné naissance, au sein de cette station de travail, à deux outils opérationnels, SDOC et HENOCH.

SDOC (Scientific DOCumentary system) est une implémentation informatique complètement paramétrable[7] de la méthode des mots associés qui permet de classer et représenter cartographiquement un ensemble de documents en se basant sur les mots-clés qui décrivent le contenu des documents. SDOC a été employé dans de nombreuses études[8] de veille menées à l'INIST dans différents domaines d'application (sciences de

---

[1] Ex-CDST Centre de Documentation Scientifique et Technique du CNRS.

[2] Cette unité a pour mission d'assurer à l'INIST une capacité d'innovation dans les technologies de l'intelligence en développant une recherche dans des domaines comme les techniques symboliques et numériques de l'intelligence artificielle appliquées à l'analyse de l'information, le traitement informatique du langage naturel en gros corpus.

[3] SERPIA : Service d'Etude et de Réalisation de Produits d'Information Avancés.

[4] Développée par Michel Callon, Jean Pierre Courtial, William Turner et Serge bauin, cf chapitre 2

[5] DRPN : Département Recherche et Produits Nouveaux.

[6] Cf section 2.4 de ce chapitre.

[7] Les possibilités de paramétrage que j'ai introduites lorsque j'ai développé ce logiciel permettent d'affiner l'interprétation des résultats. Elles sont décrites in extenso chapitre 5.

[8] Les chapitres 2,3 et 4 sont basés sur des études.

l'information, sociologie, sciences sociales, physique, etc.). Ce logiciel est également utilisé dans le cadre de recherches[9] sur le traitement automatique de la langue naturelle menées en collaboration avec l'INRIA Lorraine (Institut National de Recherche en Informatique et Automatisme).

Cette étude approfondie de la méthode des mots associés m'a permis de préciser la problématique de l'analyse de l'IST. Comment caractériser un ensemble documentaire ? Comment naviguer dans un océan d'information ? Mes travaux ont débouché sur un système permettant de coordonner l'exploitation des résultats de différentes techniques d'analyse (techniques linguistiques, classificatoires, cartographiques, etc.) appliquées à des données bibliographiques. Dénommé HENOCH[10], ce système permet de :

- rassembler et d'organiser dans un SGBD (Système de gestion de bases de données) des données bibliographiques normalisées et codifiées ainsi que les résultats de l'applications des différentes techniques d'analyse à ces données,
- distribuer ces informations sur INTERNET via une interface de navigation générée automatiquement, et adaptée à l'analyse de l'information.

HENOCH est employé régulièrement par l'INIST dans le cadre d'opérations de veille nécessitant l'analyse de gros volumes d'informations. Les bases de données hypertextes construites par HENOCH sont consultées par les partenaires de l'INIST (départements scientifiques du CNRS, centres de recherche français et étrangers, consultants, , etc.) pour produire des rapports de veille ou de tendances comme par exemple une étude sur les prions (UNIPS unité d'indicateurs de politique scientifique du CNRS), un rapport de tendance sur les plantes transgéniques[11] (Bureau Van Dijk), un rapport européen sur les thèmes clés dans le domaine des biotechnologies (rapport EUR 17342 EN, Université de Bristol Royaume Uni).

HENOCH est également un support d'enseignement de la veille technologique à l'Université de Nancy II où j'enseigne régulièrement en 2ème année d'IUT, à l'URFIST de Toulouse et de Rennes où j'ai également effectué ponctuellement des interventions[12], l'ESIEE Ecole Supérieure d'ingénieurs en Electrotechnique et Electronique de la Chambre de Commerce et d'Industrie de Paris où Xavier Polanco intervient régulièrement, l'université d'Aix-Marseille III (DEA intelligence économique), etc.

HENOCH constitue une pièce centrale dans le cadre de projets ou programmes de coopération de l'URI avec des organismes étrangers tels que le Centre de Veille technologique du Centre de Recherche Public Henri Tudor Luxembourg

---

[9] notamment le projet ILC (Ingéniérie, Linguistique et Connaissance), rapport INRIA n° 3198, juin 1997, cf section 1.2.4

[10] Henoch est le nom d'un patriarche pré-biblique qui assumait un rôle de gardien, de veilleur, d'où le nom choisi pour ce système.

[11] disponible commercialement auprès du Bureau Van Dijk (Martine Dejean), et à l'INIST.

[12] Lettre de l'URFIST de TOULOUSE n°21, juillet 1999. J'ai également effectué des présentations orales lors de séminaires ou salons où l'INIST était exposant (parmi celles-ci, je citerais, IDT 1998 journée satellite Intelligence Economique et Compétitivité, les journées IEC (Intelligence Economique et Compétitivité) 1995, 1996, 1997 organisées par SCIP FRANCE (Society of Competitive Intelligence Professionnals)

La conduite de ces deux projets pendant 10 ans m'a amené à publier régulièrement. Je me permet de fournir, page x, la liste complète de mes publications à ce jour dont voici la distribution[13] selon le type de communication :

- 8 articles dans des revues scientifiques avec comité de lecture en sciences de l'information et en informatique: Scientometrics (1997), Journal of Knowledge Organization (1995), International Journal of Scientometrics and Informetrics (1995), Solaris (1995), Hypertextes et hypermedia (1995, 1997), Génie logiciel (1991),
- 2 articles 'invités' dans le Micro-Bulletin thématique du CNRS (1997, 1999),
- 15 communications dans des congrès dont 9 articles dans des congrès internationaux avec comité de lecture et actes : International Conference of Bibliometrics, Informetrics and Scientometrics (1993, 1995, 1997), Conférence Internationale Hypertextes et Hypermedias : réalisation, outils, méthodes (1995, 1997), International Conference on Cognitive and Computer Sciences for Organizations (1993), Conférence Internationale Le Génie logiciel et ses Applications (1991), Conférence RIAO Recherche d'Informations Assistée par Ordinateur (1991), Multimedia Information Conference (1991), Les systèmes d'information élaborée (1991, 1993, 1995, 1997, 1999), Veille Scientifique et Stratégique VSST (1998).
- 1 intervention orale, en tant qu'invité lors d'un séminaire de l'ADEST (Association pour la mesure des Sciences et Techniques), le 9.12 1997 effectuant le point sur les outils de veille.

Cette activité soutenue m'amène à présenter cette thèse sous une forme originale pour la discipline : un recueil d'articles publiés. Le corps de la thèse se compose d'une sélection de 7 articles[14] illustrant chacun un aspect de la problématique de l'analyse de l'information scientifique, ainsi que deux articles en annexe retraçant la genèse de ma recherche dans ses composantes théoriques et techniques. Ce recueil est précédé par une introduction définissant cette problématique et situant mes développements dans un contexte scientifique. Il est suivi par un bilan critique et une mise en perspective de mon activité de recherche.

[13] Soit en tout 20 articles, dont 5 ont été à la fois publiés dans des actes de congrès puis sélectionnés dans des revues.

[14] Chaque article constitue un chapitre qui est précédé d'une page en couleur comportant un titre se référant au sommaire, une note de référence à l'article original et une synthèse mettant en exergue les points clés traités par rapport à la problématique.

Liste de mes publications par ordre chronologique

Les articles signalés en gras constituent le corps de cette thèse, ceux en italique figurent en annexe.

en tant qu'auteur principal

1. Grivel L. 'HENOCH, un outil d'analyse de corpus d'information scientifique et technique', Le Micro Bulletin Thématique n°3, L'information scientifique et technique et l'outil Internet, CNRS-DSI, p.27-44, 1999.

2. Grivel L., Fagherazzi H. Fourneret P. Zerouki A. 'Conception de bases de données infométriques hybrides: analyse de la pratique de trois observatoires européens et propositions', Les systèmes d'information élaborée, Ile Rousse, Corse, Edition CD-ROM (CRRM - Marseille), 1999.

3. Grivel L., Polanco X., Kaplan A. 'A computer System for Big Scientometrics at the Age of the World Wide Web', Scientometrics, vol.40, N°3, 1997, 493-506, 1997, et in proceedings of the 6th International Conference on Scientometrics and Informetrics, Jerusalem, 131-142, 1997.

4. Grivel L., Francois C., Polanco X. 'Analyse de l'information par cartographie neuromimétique et requêtes SQL sur le Web ', - 4ème Conf. Intern. Hypertextes et Hypermedias : réalisation, outils méthodes, Université Paris 8, Saint Denis, in *H2PTM97*, Editions Hermès, Vol.1, n°2, 237-248, 1997.

5. Grivel L., Polanco X., Kaplan A. 'Requêtes et navigation à partir de l'information structurée, le système HENOCH', Le Micro Bulletin, N°70, 1997, 493-506.

6. Grivel L., Mutschke P., Polanco X. 'Thematic mapping on bibliographic databases by cluster analysis : a description of SDOC environment with SOLIS', Journal of Knowledge Organization, Vol. 22, n°2, 70-77, 1995.

7. Grivel L., Francois C. 'Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique' - Solaris *n°2* "Les sciences de l'Information : Bibliométrie, Scientométrie, Infométrie", Presses universitaires de Rennes, p.81-113, 1995.

8. Grivel L., Francois C. 'Conception et développement d'un système d'information dédié à la veille scientifique basé sur les sorties des outils de classification thématique SDOC et NEURODOC' - 3ème conf. Intern. Hypertextes et Hypermedias : réalisation, outils méthodes - Editions Hermès, pp. 109-118, 1995.

9. Grivel L., Lamirel J-Ch., 'An analysis tool for scientometric studies integrated in an hypermedia environment',Proceedings of 4th International Conference on Cognitive and Computer Sciences for Organizations (ICO93), Montreal, (Quebec) Canada, pp.146-154, 1993. Et in rapport CRIN/93-R-179.

10.Grivel L., Lamirel J-Ch. 'SDOC, A Generator of Hypertext Structures', M. Feeney et S. Day (Eds), Multimedia information, Londres: Bowker Saur, p. 69-81, 1991.

**** en tant que co-auteur*

11. Polanco X., François C., Royauté J., Grivel L., Besagni D., Dejean M., Otto C., 'Organisation et gestion des connaissances en veille scientifique et technologique', **VSST'98**, Toulouse, 1998.

12. Faucompré P., Grivel L., Polanco X., Dou H., Quoniam L. 'Un lien effectif entre informations scientifiques et informations techniques', Les systèmes d'information élaborée, Ile Rousse, Corse, 1997.

13. François C., Grivel L. 'Deux éléments de la plate-forme infométrique de l'INIST : NEURODOC et HENOCH', ADEST Séminaire du 9 décembre 1997, http://www.upmf-grenoble.fr/adest/seminaires/

14. Polanco X., Royauté J., Grivel L., Courgey A. 'Infométrie et linguistique informatique, une approche linguistico-infométrique au service de la veille scientifique et technologique', Les systèmes d'information élaborée, Ile Rousse, Corse, 1995.

15. Polanco X., Grivel L., Royauté J. -'How to do things with terms in informetrics: terminological variation and stabilization as science watch indicators'- Proceedings of the 5th International Conference of the International Society for Scientometrics and Informetrics -, Chicago, Illinois, pp.435-444, 1995.

16. Polanco X., Grivel L. -'Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United State of America)', JISSI International Journal of Scientometrics and Informetrics, Vol.1, Nr. 2, june 1995, pp 123-137. 4th International conference of Bibliometrics, Informetrics and Scientometrics -, Berlin, Germany. 1993

17. Polanco X., Grivel L., François C., Besagni D. "L'infométrie, un programme de recherche", Journées d'études sur les systèmes d'information élaborée de la SFBA, Ile Rousse, Corse, Document n° 3 des Actes, 9p, 1993.

18. Ducloy J., Charpentier P., François C., Grivel L. 'Une boîte à outils pour le traitement de l'Information Scientifique et Technique', 4es. Journées Internationales Le Génie logiciel et ses applications. Toulouse, 9-13 Décembre 1991, p. 239-254 ; et dans Génie logiciel, n° 25, p. 80-90, 1991.

19. Ducloy J., Grivel L., Lamirel J-Ch., Polanco X., Schmitt L., 'INIST's Experience in Hyper-Document Building from Bibliographic Data Bases'. Proceedings of RIAO 91 Conference - Barcelone (Spain), vol. 1, 1991.

20. Polanco X., Schmitt L., Besagni D., Grivel L. 'A la recherche de la diversité perdue : est-il possible de mettre en évidence les éléments hétérogènes d'un front de recherche ?', les systèmes d'information élaborée, Ile Rousse, Corse, p. 273-292, 1991.

Chapitre 1

De l'analyse de l'information scientifique à l'hypertexte

Ce chapitre définit ma problématique de recherche : l'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'Information Scientifique et Technique (IST), positionne mes développements dans un contexte scientifique interne (l'Unité Recherche et Innovation à l'INIST) et externe (en France et à l'étranger). Il introduit également les chapitres suivants.

La première partie situe l'analyse de l'IST au sein d'une discipline (l'infométrie) et en propose une définition opérationnelle. L'analyse de l'information présente un fort caractère exploratoire. Si l'on se fixe comme objectif de faire émerger (découvrir) automatiquement la structure cognitive d'un grand ensemble de documents sans passer par un plan de classement pré-établi, les technologies de classification automatique et de représentation graphique (cartes) développées en analyse de données sont les plus adaptées. Si de plus, on se propose de représenter les connaissances véhiculées par les textes scientifiques et techniques sous leur forme écrite, il est indispensable de s'appuyer sur des techniques linguistiques. Dans ce cadre, l'analyse de l'IST peut alors être définie comme l'application de techniques de traitement automatique du langage naturel, de classification automatique et de représentation graphique (cartographie) du contenu cognitif et factuel des données bibliographiques.

La deuxième partie explicite les liens entre le concept d'hypertexte et les méthodes d'analyse. D'une métaphore, la navigation dans un océan d'information se déduit un principe de conception qui est aujourd'hui commun à un certain nombre d'équipes de recherche dans notre domaine d'application (l'analyse de l'IST) : générer automatiquement des hypertextes avec leur carte de navigation. Ce principe se concrétise sur le plan opérationnel par un système générateur d'hypertextes accompagnés de leur carte de navigation : le système HENOCH. Ce système trouve sa place au sein de la plate-forme infométrique de l'URI qui est décrite dans son ensemble.

1. L'analyse de l'information scientifique et technique (IST)

Cette section est une réactualisation de l'article fondateur du Programme de Recherche en Infométrie figurant en annexe 1. Elle pose la problématique de l'analyse de l'IST, la situe au sein d'une discipline (l'infométrie) et enfin propose une définition opérationnelle de l'analyse de l'IST.

1.1 La problématique de l'analyse de l'IST et son intérêt pour un institut tel que l'INIST

L'accroissement de l'activité scientifique jointe à l'éclosion des nouvelles technologies de l'information se traduisent par une croissance remarquable de l'information scientifique et technique (IST) disponible sous forme électronique. L'information scientifique et technique est produite en abondance, archivée quasi systématiquement (banques de données documentaires, documentation technique), signalée (bases de données bibliographiques, bases de données brevets) et diffusée (CDROM, Internet, serveurs en ligne) sous forme électronique. Ainsi en France, L'Institut de l'Information Scientifique et technique (INIST), au sein du Centre National de la Recherche Scientifique (CNRS), a pour mission de collecter, traiter et diffuser les résultats de la recherche scientifique et technique internationale en France et à l'étranger. Le fonds documentaire de l'INIST couvre la plus grande partie de la recherche scientifique et technique mondiale :les publications en série (27 OOO titres de périodiques internationaux dont 9 000 environ correspondent à des abonnements en cours en 1999), la littérature grise (plusieurs centaines de milliers de documents échappant aux circuits commerciaux traditionnels de l'édition comme par exemple, les thèses , les comptes de congrès ou les rapports scientifiques). Pour donner un ordre de grandeur, l'INIST fournit actuellement 700 000 copies de documents par an. En 1999, plus de 6 000 périodiques sont analysés pour alimenter deux bases bibliographiques multilingues et multidisciplinaires, PASCAL en sciences, technologie et médecine, FRANCIS en sciences humaines et sociales, rassemblant respectivement 13 millions et 2,3 millions de références.

Dès lors que la description bibliographique de ces millions de documents se trouve accessible en ligne sous format électronique, il est possible d'utiliser cette mine d'informations à des fins stratégiques : la gestion stratégique de la recherche et de la technologie, la détermination d'une politique scientifique, la surveillance d'un domaine scientifique avec ses multiples aspects (la synthèse d'information, l'observation des tendances, le positionnement des acteurs de la recherche, etc.). La valorisation de l'information, et notamment de ses bases bibliographiques, faisant partie des missions de l'INIST, l'institut s'est doté d'une unité de recherche (URI Unité Recherche et Innovation), ayant pour but principal la conception et la production d'instruments -c'est-à-dire d'indicateurs, de méthodes et d'outils- d'analyse de l'information scientifique et technique (IST). La section suivante positionne l'analyse de l'IST par rapport à l'infométrie, discipline dont l'objet d'étude est précisément l'information.

1.2 L'infométrie : discipline carrefour pour l'analyse de l'IST

L'analyse de l'information scientifique et technique telle qu'elle est signalée dans les bases de données fait l'objet de recherches dans le cadre d'une discipline carrefour dans ce type d'activité: l'infométrie.

Le terme infométrie désigne l'ensemble des activités métriques ayant comme objet

l'information, couvrant aussi bien la scientométrie que la bibliométrie[1]. On trouvera dans [POLANCO 1995] une description de la genèse de ces techniques et notamment de la théorie du développement de la science due à Derek de Sola Price dans les années 60.

Les deux grands champs d'application de l'infométrie sont d'une part l'évaluation de la recherche à travers sa production scientifique [COURTIAL 1990], [ZITT 1996] et d'autre part la veille scientifique et technique [QUONIAM 1988, 1992], [KISTER et al. 1993], [DOU 1995] définie comme "*l'observation et l'analyse de l'environnement scientifique, technique, technologique et des impacts économiques présents et futurs, pour en déduire les menaces et les opportunités de développement*"[2]. Dans les deux cas, les bases de données bibliographiques sont une source importante d'information [JAKOBIAK 1996, LECOADIC 1994].

L'infométrie fournit en effet des outils et des méthodes pour traiter de grandes quantités d'information. Ces méthodes vont de la statistique descriptive aux analyses multidimensionnelles, en passant par des techniques de classification et de cartographie. Elles s'appuient sur des lois (Lotka, Zipf, Bradford), appelées également lois bibliométriques ou infométriques, qui sont à l'analyse de l'IST ce que la loi de Pareto (individus-revenus) est à l'économie, ou les statistiques démographiques à l'étude des populations. Ces trois lois aident à déterminer des critères quantitatifs, pour sélectionner **un ensemble représentatif** (d'un point de vue analyse de données) d'auteurs [LOTKA 1927], de périodiques [BRADFORD 1934) ou de mots-clés [ZIPF 1949], sur lesquels on peut ensuite appliquer des techniques d'analyse descriptive ou d'analyse de données pour construire des indicateurs de politique scientifique. Signalons cependant que le caractère hyperbolique de ces distributions interdit l'application de toute une famille de méthodes statistiques fondée sur une distribution de type répartition normale (moyenne, écart type, etc.), ce qui limite dans une certaine mesure leur emploi [LAFOUGE 1998]. [ROSTAING 1996] et [POLANCO 1995] constituent une introduction accessible et documentée, en français, à ces méthodes.

La manière d'aborder l'analyse infométrique d'un corpus de données diffère selon l'objectif à atteindre : la veille ou l'évaluation (la mesure) de l'activité de recherche. Cette distinction opère sur deux critères : la couverture du sujet d'étude et la réactivité. Dans le cas de la veille la couverture tente d'être exhaustive sur le sujet, alors que dans le cas de l'infométrie d'indicateurs, elle se doit d'être représentative. Pour citer D. J de Solla Price, dans le cas de la veille, on veut tout savoir sur 'Georges', pour des raisons diverses (mon futur employeur, mon concurrent, etc.), alors que dans le cas de la mesure de l'activité de recherche, on ne s'intéresse pas au cas de 'Georges' en particulier, les indicateurs sont fondés sur une logique de comparaison. Dans ce cadre, la sélectivité est préférable à l'exhaustivité, c'est à dire la recherche d'une couverture (le plus souvent un ensemble de revues cœur du domaine) répondant à des critères qualitatifs les plus clairs et contrôlables possibles. Le critère le plus utilisé est le prestige de la revue fondé sur le calcul des citations reçues. Ce qui, par le jeu des avantages cumulés induit une auto-validation de la couverture. Ce critère est le plus souvent complété par l'avis d'un comité d'experts. L'aspect calcul et type d'indicateurs dans le cadre de l'évaluation de la production scientifique et technique à travers la littérature scientifique est abordé de

---

[1] Cf. définitions complètes relatives à l'infométrie, la bibliométrie et la scientométrie en annexe 1.

[2] François JAKOBIAK. *Exemples commentés de veille technologique*. Paris : Les Editions d'Organisation, 1992, p. 27.

manière très fouillée dans [MOED 1996], [GLANZEL 1996].

Veille et évaluation de la recherche se distinguent également au niveau du degré de réactivité attendu. La veille privilégie la réactivité au prix éventuellement d'une information bruitée, tandis que dans le cas de l'évaluation de la recherche, le rythme de production de rapports est le plus souvent annuel. Le temps passé à nettoyer les données peut être plus long (constitution et utilisation de tables de nomenclatures [GRIVEL 1999], chapitre 7), car les résultats attendus doivent être les plus robustes possibles (au sens statistique du terme).

L'évaluation de la recherche et la veille scientifique et technique définissent le contexte social dans lequel se situe l'analyse de L'IST, tandis que l'infométrie définit son champ disciplinaire. Il reste maintenant à définir l'analyse de l'information sur le plan opérationnel.

1.3 Une définition opérationnelle de l'analyse de l'IST

L'IST, telle qu'elle est signalée dans les bases bibliographiques, se présente sous forme textuelle et structurée. La sémantique est exprimée par les étiquettes décrivant les champs (titres, auteurs, affiliations, date de publication, etc.), et éventuellement par l'ordre des données. La phase de traitement de l'information a pour objectif de rendre exploitables les informations traitées. D'un point de vue infométrique, l'analyse de l'IST a pour objectif de caractériser un ensemble documentaire sur le plan cognitif et factuel ('qui fait quoi, où, collabore avec qui, quand ?'). La sortie attendue est une présentation de l'information, non pas sous sa forme brute mais sous une forme élaborée (classée, structurée), de façon à ce que l'usager puisse en dégager le sens ou les aspects stratégiques.

Il existe essentiellement deux approches [LEBART 1988]:

- classer les documents en les affectant à des classes préexistantes. Par exemple, en utilisant l'analyse discriminante.
- classifier les documents, c'est à dire les regrouper (découvrir les classes) à partir de mesures de similarité. La classification hiérarchique, la méthode des nuées dynamiques font partie des techniques de classification couramment utilisées.

L'analyse de l'information présente un fort caractère exploratoire. Si l'on se fixe comme objectif de faire émerger (découvrir) automatiquement la structure cognitive et factuelle d'un grand ensemble de documents sans passer par un plan de classement pré-établi, les technologies de classification automatique et de représentation graphique (cartes) développées en analyse de données sont les plus adaptées.

Si de plus, on se propose de représenter les connaissances véhiculées par les textes scientifiques et techniques sous leur forme écrite, il est indispensable de s'appuyer sur des techniques linguistiques [POLANCO 1996].

Dans ce cadre, l'analyse de l'IST peut alors être définie comme l'application de techniques de traitement automatique du langage naturel, de classification automatique et de représentation graphique (cartographie) du contenu cognitif et factuel des données bibliographiques.

2 L'hypertexte et les méthodes d'analyse de l'IST

Cette sous-section explicite les liens entre le concept d'hypertexte et les méthodes d'analyse. D'une métaphore, la navigation dans un océan d'information (section 2.1), se

déduit un principe de conception (section 2.2), qui est aujourd'hui commun à un certain nombre d'équipes de recherche (section 2.3) dans notre domaine d'application (l'analyse de l'IST) : générer automatiquement des hypertextes avec leur carte de navigation. Ce principe se concrétise sur le plan opérationnel par un système générateur d'hypertextes accompagnés de leur carte de navigation: le système HENOCH, au sein de la plate-forme infométrique de l'URI (section 2.4).

2.1 Naviguer dans un océan d'information

Le point de départ de mon travail été fondé par la constatation suivante. En 1990, époque où j'ai débuté mes travaux, les outils d'analyse de l'information étaient déjà relativement nombreux et variés du point de vue des méthodes mises en œuvre [COURTIAL 1990] mais l'exploitation et l'interprétation des résultats obtenus restaient mal aisées. Sans doute parce que le processus d'analyse de l'information est un mélange d'exploration informelle intuitive (par association d'idées) et d'exploitation méthodique de l'information élaborée par différents outils d'analyse et que les outils développés à cet époque ne prenaient pas en compte suffisamment cet aspect. Ceci suppose d'assister le travail d'interprétation des sorties des méthodes d'analyse de l'information en favorisant les interactions entre les schémas mentaux de l'utilisateur (sa représentation du domaine couvert par la littérature scientifique) et différentes représentations cognitives fournies par les méthodes d'analyse.

L'hypothèse effectuée dans mes recherches est que ces techniques d'analyses devaient être coordonnées par une métaphore, également exprimée par [LELU 1993]: la navigation dans un océan d'information. Pour s'y retrouver, avoir une vue d'ensemble, se positionner et positionner ses concurrents, l'usager doit disposer d'une carte du domaine, d'une "boussole" pour orienter sa carte (sa connaissance du domaine) et de méthodes d'analyse pour faire le point, connaître son positionnement (se situer par rapport aux représentations fournies par les méthodes d'analyse) et celui des autres.

L'hypertexte, en tant que principe d'organisation de l'information, semble[3] le moyen le plus adéquat pour modéliser cette organisation, et, en tant que technologie, mettre en place concrètement les mécanismes d'exploration et les interactions nécessaires à l'interprétation des résultats d'analyse.

2.2 La génération automatique d'hypertexte et les techniques d'analyse

Il n'est pas de mon propos de faire un historique [SERRES 95] sur l'hypertexte, dont l'usage, avec l'essor d'internet, s'est largement popularisé, mais plutôt d'introduire l'hypertexte sur le plan conceptuel et technique.

La définition ci-dessous pour le terme hypertexte est suffisamment générale pour s'appliquer à tout type de document et pas seulement aux documents textuels. Un hypertexte est un ensemble d'unités d'information (« noeuds »), qu'un utilisateur peut parcourir de façon informelle libre et exploratoire au moyen de liens proposés par le système. Les hypertextes ont pour vocation d'articuler et d'organiser des entités plus au moins atomiques d'informations, à l'aide de relations existant entre ces granules de connaissance.

---

[3] Les études effectuées dans les chapitres 2, 3 et 4 ainsi que le chapitre 8 corroborent cet avis. Voir également dans [LEVY 1990], 'la métaphore de l'hypertexte' (chapitre 1) pour une analyse des principes de l'hypertexte et le besoin de cartes interactives pour naviguer.

Dans notre problématique, ces entités ou « noeuds » peuvent être des documents, des auteurs, des revues, des agrégats (clusters) de documents ou de mots-clés, des indicateurs, des cartes, etc. Ces noeuds peuvent être édités ou calculés. Les relations existant entre ces entités constituent les liens hypertextuels qui peuvent être de deux types : liens de références, liens hiérarchiques. Ces liens peuvent être établis manuellement ou calculés automatiquement.

Dès lors qu'il s'agit d'analyser de gros volumes d'information, il n'est plus question de construire l'hypertexte manuellement mais de le générer, c'est à dire de calculer dynamiquement les noeuds et liens qui constituent l'hypertexte à partir de textes ou des données déjà disponibles. C'est là que peuvent êtres mises à profit certaines des techniques citées en section 1.3. Mais pour éviter la désorientation de l'usager devant l'énorme quantité de liens générés automatiquement, une représentation cartographique de l'ensemble du contenu de la base est nécessaire. L'enjeu est alors de générer automatiquement ces hypertextes avec leur carte de navigation.

La génération automatique de noeuds et de liens hypertextes utilise trois approches complémentaires [BALPE 1995]:

1) une approche structurelle : une donnée bibliographique, par exemple, est structurée, découpée en unités élémentaires hiérarchisées, avec des renvois multiples (bibliographie, notes, liens entre auteurs et affiliation, etc.) qui peuvent être utilisés pour générer des liens. SGML[4], et son évolution, XML[5], sont les normes utilisées aujourd'hui pour décrire la structure logique de documents.

2) une approche linguistique : cette approche consiste à considérer la langue du texte comme porteuse d'informations analysables pour en extraire des liens hypertextes. En récupérant toutes les informations que peut fournir le texte d'un document, il est possible d'en extraire un ou plusieurs réseaux de parcours possible. Le principe consiste à exhiber par des moyens automatiques une organisation à partir des éléments d'information (unités textuelles élémentaires ou termes) qu'il est possible d'extraire du corpus, c'est à dire lier ces éléments entre eux (réseaux de sens de type encyclopédique, comme par exemple dans le système TAIGA [MARTEAU 95], liens de cooccurrence comme dans le système SAMPLER [JOUVE 1998] (issus de la méthode des mots associés [CALLON et al. 1983, 1986, 1993], [MICHELET 1988]), liens de variations flexionnelles ou syntaxiques d'un terme complexe tel que le groupe nominal [ROYAUTE 1999] (section 2.4.3 et chapitre 4,), etc.

3) une approche statistique : cette approche consiste à considérer une collection de documents pour en extraire des caractéristiques. Elle permet de structurer l'information en distinguant dans un premier temps leurs possibilités de regroupements, au sein d'une entité de niveau supérieur (une classe), d'entités similaires du point de vue des caractéristiques extraites, et, dans un deuxième temps,

---

[4] SGML, Standard Generalized MarkUp Language, meta-langage permettant de construire des langages de balisage de documents pour rendre compte de leur structure logique.

[5] XML (eXtensible Markup Language) est une version modernisée et simplifiée de SGML, issue des travaux du W3C. XML retient les caractéristiques essentielles de SGML en l'épurant de ses caractéristiques les plus complexes à mettre en œuvre et en apportant de puissants de mécanismes de liens, étendant ceux présents dans HTML. Il existe une traduction en français de la norme XML, http://babel.alis.com/web_ml/xml

la création de cartographies de l'ensemble de ces classes en les situant les unes par rapport aux autres [TEIL 1991], [LELU 1993], [SMALL 1997, 1999].

Sur la base de ces principes, un environnement d'analyse de l'IST devrait comporter non seulement un ensemble d'outils d'analyse disponibles au sein d'une plate-forme, mais également un 'observatoire', véritable système d'information que nous appelons base infométrique, où l'utilisateur peut stocker, explorer et exploiter méthodiquement, selon la métaphore 'navigationnelle' décrite en section 2.1, les résultats quantitatifs ou qualitatifs de l'application de différentes méthodes d'analyse sur des données relatives à une problématique particulière.

Le développement d'un tel environnement d'analyse (section 2.4) est l'un des buts de l'unité de recherche de l'INIST (URI Unité Recherche et Innovation), but que nous partageons avec un certain nombre d'équipes en France et à l'étranger.

2.3 Contexte scientifique

Un certain nombre d'équipes en France et à l'étranger partagent ce point de vue, à savoir qu'il est nécessaire, notamment dans notre domaine d'application, de générer automatiquement les hypertextes avec leur carte de navigation. Dans le cadre de mon travail, j'ai effectué un suivi des équipes travaillant sur le sujet. Ces équipes, à l'instar de l'URI, mêlent le plus souvent des chercheurs en analyse de données, analyse linguistique et informatique. Ni exhaustif, ni comparatif, le tableau des équipes ci-dessous décrit le nom du ou des logiciels développés, le thème de recherche et fournit quelques références. Pour une étude comparative de différents logiciels de veille intégrant certaines des techniques décrites plus haut, voir [ROUSSEAU 98].

Département Hypermedia UFR 6 Université Paris VIII		
NEURONAV +	hypertexte dynamique et extraction terminologique, classification neuronale et cartographie	http://hypermedia.univ-paris8.fr/ [Lelu et al. 1997 et 1998],
Département Informatique des Images, des Sons et des Textes, IRIT Institut de Recherche en Informatique de Toulouse		
TETRALOGIE	Exploration dans les bases d'informations et découverte de connaissances, extraction terminologique, méthodes factorielles	http://atlas.irit.fr http://www.irit.fr/SSI/ACTIVITES/EQ_SIG/themes/datamining/exploration.html [Dkaki et al. 1997 et 1998]
ECAM European Centre for Applied Mathematics, IBM, Paris		
TKS (Text Knowledge Server) Technology Watch	fouille de données textuelles veille technologique, extraction terminologique et analyse relationnelle	http://www.fr.ibm.com/france/ecam/soluttm.htm [MARCOTORCHINO 1991], [HUOT 1992]
CRRM Centre de Recherche Rétrospective de Marseille, Université d'Aix Marseille III		
DATAVIEW,	infométrie appliquée à la	http://crrm.univ-mrs.fr/

DATABLOCK MATRISME	veille technologique , Internet et analyse réseau	[BOUTIN et al. 1998] [LEVEILLE et al. 1998]
CWTS Centre for Science and Technology Studies, Leiden University (Hollande)		
	infométrie, évaluation de la recherche et systèmes d'informations et analyse mots associés	http://sahara.fsw.leidenuniv.nl/cwts/noframes/cwtshome.html [Noyons, AFJ Van Raan 1998]
ISI Institute for Scientific Information (USA)		
SCI-VIZ (prototype)	infométrie et systèmes d'informations et cartographie de la science	http://www.isinet.com [SMALL 1997 et 1999]
Austrian Research Center Seibersdorf, Department Technology Studies (Autriche)		
	infométrie, et cartographie de la science	[KOPCSA et SCHIEBEL 1998]
School of Library and Information Science, University of Wisconsin-Milwaukee, (USA)		
Hyperlinx	infométrie et hypertexte	[WOLFRAM 1996]
Neural Networks Research Centre, Helsinki University of Technology (Finlande)		
Websom	cartographie, réseaux neuronaux et hypertexte	http://websom.hut.fi/ [KOHONEN et al. 1995]

Tableau 1 : contexte scientifique

2.4 La plate-forme infométrique de l'URI pour analyser l'IST

L'URI a pour but principal, la conception et la production d'instruments [c'est-à-dire d'indicateurs, de méthodes et d'outils] d'analyse de l'information scientifique et technique (IST). Cette activité se traduit sur le plan informatique par le développement d'une plate-forme logicielle. La plate-forme infométrique (Figure 1) est le nom générique donné à l'ensemble des outils de l'URI. Elle intègre une certain nombre de techniques :

1. des techniques linguistiques fournissant des mécanismes d'extraction terminologique sur du texte intégral en anglais et en français qui permettent de s'affranchir de l'indexation manuelle [ROYAUTE 99]. Ces techniques sont intégrées au sein d'une plate-forme d'ingénierie linguistique dénommée ILC.

2. des statistiques descriptives fondées sur les distributions bibliométriques,

3. des techniques de classification hiérarchique et non hiérarchique et de cartographie (ACP, diagramme stratégique, réseaux neuronaux) pour la structuration de l'information. Ces techniques sont intégrées dans deux programmes, SDOC [GRIVEL 1995a] et NEURODOC [LELU 1993], [FRANCOIS 1998], etc.

4. des techniques d'ingénierie documentaire basées sur l'emploi de SGML[6] [DUCLOY et al. 91], d'un SGBD relationnel et d'un serveur Web intégrés au sein du

[6] SGML, Standard Generalised Mark Up Language, norme [ISO 8879], [GOLDFARB 90], HERWIJNEN 90], Le format SGML (Standard Generalized Markup Language)

logiciel HENOCH [GRIVEL et FRANCOIS1995b] , [GRIVEL et al. 1997], [GRIVEL 1999].

Une chaîne de constitution de corpus et de traitements s'appuyant sur cette plate-forme a été mise en place. Le traitement se décline en 5 phases successives:

- reformatage des notices selon la norme SGML,
- traitement statistique portant sur les éléments bibliographiques des notices (auteurs, périodiques, dates, indexation), le programme MIRIAD[7],
- traitement linguistique d'acquisition terminologique (la plate-forme ILC[8]),
- traitement de classification et de cartographie par les logiciels SDOC[9] ou NEURODOC[10], et enfin
- stockage par le logiciel HENOCH[11] des résultats de ces traitements antérieurs et mise à disposition sur le Web selon une interface basée sur la métaphore décrite en section 2.1.

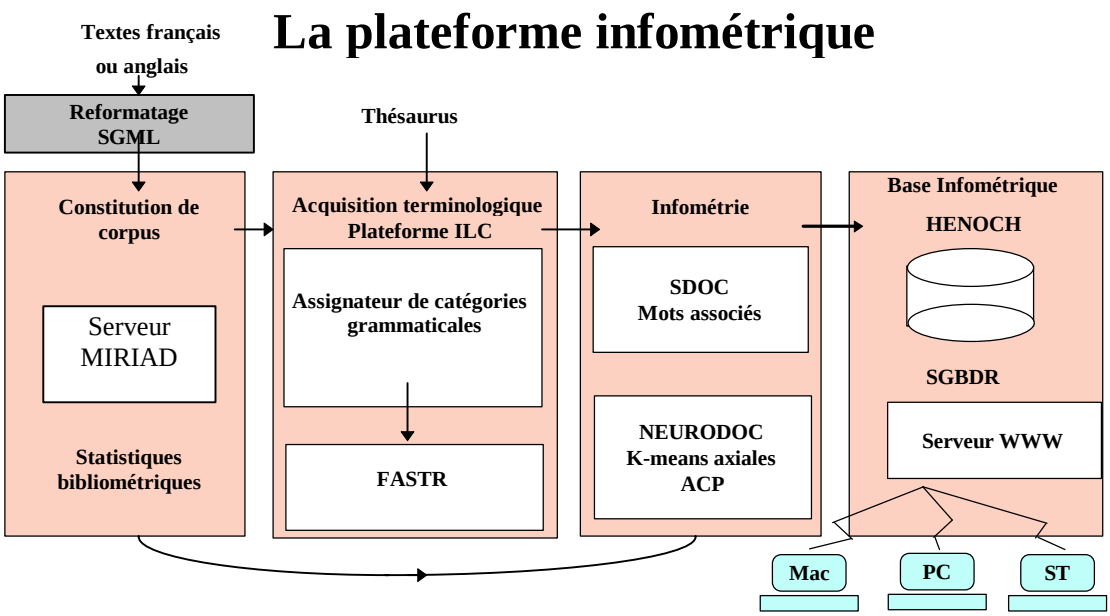


Figure 1[12] : la plate-forme infométrique

2.4.1 Reformatage

---

donne des règles de balisage pour décrire des structures arborescentes où chaque noeud est identifié par une étiquette. Baliser un document consiste à insérer dans le texte des chaînes de caractères qui donnent de l'information sur le contenu du document.

[7] Ce programme a été réalisé par D. Besagni (URI).

[8] La plate-forme ILC a été réalisée par Jean Royauté (URI) et Chantal Muller qui a quitté l'URI depuis.

[9] Ce programme a été réalisé par L. Grivel (URI).

[10] Ce programme a été réalisé par C. François (URI).

[11] Ce système a été conçu par L. GRIVEL (URI) puis réalisé en collaboration avec 3 ingénieurs en informatique (A. Kaplan, B. Levy, Ch. Broussaudier) de l'ESIAL.

[12] Figure extraite du document technique 'Réalisation d'une station de l'analyse de l'information', INIST, 14.01.1998.

La nature textuelle des données à analyser, la diversité de leur structure, le nombre de champs différents à traiter pour mener à bien une étude infométrique, nous ont amenés à adopter la norme SGML[13] pour la description de la structure logique de tous les document manipulés par les outils de la plate-forme. La première étape est donc un reformatage des notices bibliographiques afin de les rendre exploitables tout au long de la chaîne. Une fois que toutes les données sont homogénéisées dans un format pivot unique, il est plus facile de concevoir des outils génériques. La plupart des traitements sur de tels documents se réduisent à associer des actions à une balise. Ces caractéristiques nous ont conduits à développer une boite à outils (appelée ILIB) basée sur SGML et sur les outils du système d'exploitation UNIX (cf. Annexe II [DUCLOY *et al* 1991]).

2.4.2 Statistique descriptives

MIRIAD comprend plusieurs modules permettant de faire des statistiques standard sur différents formats de notices, à commencer par ceux utilisés à l'INIST pour les bases PASCAL et FRANCIS. Ces statistiques portent :

- sur les éléments bibliographiques des notices : distribution par code de classement, par type de document, par date de publication, par langue, par affiliation des auteurs et par pays de publication ;
- sur les périodiques : nombre de notices produites par titre et nombre de titres par pays de publication ;
- sur l'indexation : distribution des mots-clés par notices et distribution des mots-clés par fréquence.

Leur emploi dans le cadre de l'analyse de l'IST est décrit dans le chapitre 2.

MIRIAD permet de représenter sous forme de tableaux ou histogrammes, la distribution des éléments bibliographiques dont l'interprétation est fondée sur les lois bibliométriques (Lotka, Zipf et Bradford). Les résultats peuvent être fournis en format HTML par FTP ou sur une disquette, et sont donc visualisables à l'aide d'un outil de navigation WWW.

MIRIAD comprend aussi un serveur interne à l'INIST dont le but est de réaliser des statistiques à la demande sur les notices issues des bases PASCAL et FRANCIS. Par le biais d'un outil de navigation WWW, les utilisateurs peuvent constituer un corpus de notices à partir de requêtes booléennes, définir et lancer une analyse statistique.

La recherche se fait sur l'ensemble de la base PASCAL depuis 1992, avec une mise à jour hebdomadaire. ***MIRIAD*** fournit aux utilisateurs un compte-rendu détaillé du résultat de leur recherche et permet la visualisation des notices obtenues, donnant en cela la possibilité de vérifier la pertinence de leurs requêtes et de les modifier si besoin est.

2.4.3 La plate-forme d'ingénierie linguistique ILC

La plate-forme ILC permet d'indexer une collection de documents (corpus) par la reconnaissance de termes présents à la fois dans un lexique terminologique ou un

[13] SGML : Standard Generalized Mark-up Language.

thésaurus et dans le corpus. Les traitements terminologiques réalisés s'appuient sur l'intégration de deux principaux outils linguistiques : l'analyseur FASTR [JACQUEMIN 1994], un outil linguistique de traitement du groupe nominal et l'assignateur de catégories grammaticales [ROYAUTE et JACQUEMIN 1993]. Ce dernier réalise l'étiquetage des termes du lexique utilisé. Chacun des mots du terme est identifié par son lemme (racine du mot), sa catégorie flexionnelle qui permet d'identifier un nom avec ses pluriels réguliers et irréguliers, une catégorie syntaxique (par exemple verbe nominalisé), son genre (masculin, féminin).

L'analyseur FASTR permet le repérage des termes et de leurs variantes. Un ensemble de meta-règles (qui varient selon la langue) opère sur les termes étiquetés et définit les possibilités de variations flexionnelles et syntaxiques des termes, ce qui rend possible leur identification sous des formes qui peuvent être éloignées de la forme enregistrée dans le lexique ou le thésaurus.

ILC fonctionne actuellement sur le français et l'anglais. Les termes collectés, anglais ou français, peuvent donc être reconnus sous leurs formes d'enregistrement dans le lexique terminologique de départ, sous les formes singuliers ou pluriels (variations flexionnelles) ou sous des formes syntaxiques variantes [ROYAUTE 1999].

Trois sortes de variations syntaxiques sont traitées :

(a) la *variation d'insertion* concerne tout mot à l'intérieur du groupe nominal, à l'exception de la plupart des mots grammaticaux. Par exemple, *X ray absorption spectroscopy* est associé au terme *X ray spectroscopy ;*

(b) la *variation de coordination* concerne toute forme coordonnées de mots (adjectifs ou noms) à l'intérieur du groupe nominal. Par exemple, *differential and integrated cross sections* est associé au terme *Differential cross section* ;

(c) la *variation de permutation* implique tous les mots ou les groupes de mots pouvant permuter autour d'un élément pivot (prépositions ou séquences verbales). Par exemple, *range of power modulation frequency* est associé au terme *Frequency range*.

Leur emploi dans un contexte d'analyse de l'information est décrit dans le chapitre 3.

2.4.4 Les outils de classification et cartographie : SDOC et NEURODOC

Ces deux outils utilisent les mots-clés qui indexent les références bibliographiques pour mettre en évidence des structures thématiques, indicateurs de centres d'intérêt ou thèmes. Ceux-ci sont ensuite disposés sur un espace à 2 dimensions appelé "carte thématique".

a)SDOC

SDOC est un ensemble de modules implémentant la méthode des mots associés [CALLON et al. 1983, 1986], [MICHELET 1988]. Initialement orientée au service d'une analyse des sciences et techniques dans un cadre sociologique, cette méthode est ici utilisée dans un cadre Science de l'Information au service de l'analyse de l'information scientifique et technique. La méthode est basée sur la cooccurrence des mots-clés pour mettre en évidence la structure de leurs relations (réseaux lexicaux). La notion de cooccurrence est essentielle. En effet, si on considère que deux documents sont proches

parce qu'ils sont indexés par des mots-clés similaires, alors deux mots-clés figurant ensemble dans un grand nombre de documents seront considérés comme proches.

L'emploi d'un indice statistique permet de normaliser la mesure de l'association entre deux mots-clés. L'indice utilisé est l'indice d'équivalence : la cooccurrence au carré des mots-clés i et j, divisée par le produit de leurs fréquences respectives. Les valeurs varient entre 0 et 1. Cet indice est analogue aux indices bien connus de Dice, de Jaccard et de Salton.

Ensuite, ***SDOC*** applique un algorithme de *classification ascendante hiérarchique* (CAH) dit *du simple lien* (*single link clustering*), afin de construire des classes ou clusters de mots proches les uns des autres n'excédant pas une taille maximale. Un cluster est donc constitué de mots associés les uns aux autres (*associations internes ou associations intra-cluster*). Les clusters peuvent avoir des relations entre eux (*associations externes ou associations inter-cluster*).

Après le processus de classification des mots-clés, les documents sont affectés aux clusters en fonction de leur indexation.

Les clusters sont ensuite positionnés sur un plan bidimensionnel (Y, X) selon leur "densité" et "centralité", constituant ainsi une carte :

- la densité (Y) d'un cluster est exprimée par la valeur moyenne des associations entre mots-clés formant le cluster, ou associations internes ;
- la centralité (X) d'un cluster est exprimée par la valeur moyenne des associations entre les mots qui le constituent et les mots d'autres clusters, ou associations externes.

Sur une telle carte, la proximité entre deux clusters indique qu'ils sont structurellement proches, mais ne présage pas de leur proximité sémantique. Les cartes ne sont pas seulement un moyen de visualisation, elles représentent aussi une méthode d'analyse dans la mesure où elles permettent d'évaluer la position des thèmes entre eux dans un espace géométrique de représentation.

SDOC est complètement paramétrable, c'est à dire qu'il est possible de définir le nombre maximal de mots-clés composant un cluster, de limiter le nombre d'associations inter et intra clusters, de faire des filtrages sur la fréquence des mots-clés, sur le nombre de cooccurrences, sur le nombre de documents composant le cluster, etc. L'intérêt de ces possibilités de paramétrage est décrit plus spécialement dans les chapitres 4 et 5.

b) NEURODOC[14]

NEURODOC est un ensemble de modules implémentant la méthode de K-means axiales [LELU 1993], un algorithme de classification non hiérarchique et une analyse en composantes principales (ACP) pour une représentation des classes obtenues sur une carte.

[14] NEURODOC, s'intègre aujourdh'ui dans une famille d'outils basés sur des réseaux neuronaux développées à l'URI [POLANCO et al. 1997, 1998]

A partir d'une représentation vectorielle des données, la méthode des k-means axiales considère l'ensemble des documents comme un nuage de points plongé dans un espace géométrique où chaque dimension correspond à un mot-clé. Elle est caractérisée par une représentation des classes par des vecteurs pointant vers les zones de forte densité du nuage. Tandis que les techniques de classification non hiérarchiques usuelles représentent les k classes recherchées par leur centre de gravité, la méthode k-means axiales définit les k classes par k demi-axes passant par l'origine de l'espace géométrique, ou k vecteurs unitaires pointant dans la direction des ces demi-axes. Cette méthode, paramétrée par le nombre maximal de classes désiré (k) et le seuil d'appartenance des documents et des mots-clés dans les classes, permet de construire des classes d'un type particulier :

- ces classes sont recouvrantes car un document ou un mot-clé peut appartenir à plusieurs classes à la fois ;
- les éléments, documents et mots-clés de chaque classe, sont ordonnés selon leur degré de ressemblance au type idéal de la classe.

Afin de positionner les classes obtenues les unes par rapport aux autres sur une carte, l'ensemble des classes est traité comme un nuage de points. Une ACP recherche les directions d'allongement maximum de ce nuage permettant de déterminer un plan sur lequel tous les points sont ensuite projetés orthogonalement.

Les classes obtenues sont des indicateurs des thèmes ou des centres d'intérêt autour desquels s'agrège l'information, tandis que la carte propose une visualisation globale des thèmes et représente un indicateur stratégique permettant d'apprécier la position relative des classes dans l'espace de connaissance.

Les deux outils SDOC et NEURODOC sont décrits plus précisément et comparés dans le chapitre 5. Ils peuvent traiter aussi bien des textes indexés manuellement ou par la plate-forme ILC.

2.4.5 La génération automatique d'hypertextes dynamiques sur le Web : HENOCH

HENOCH est un générateur d'applications hypertextes avec carte de navigation. Il établit une passerelle entre un système producteur d'indicateurs infométriques, un système de gestion de bases de données (SGBD) relationnel, et un navigateur sur le Web. HENOCH permet de stocker les résultats des traitements infométriques linguistiques et statistiques au sein d'une base de données ORACLE puis de distribuer ces résultats sur le Web.

Sur le plan informatique, le système HENOCH assure deux fonctions principales :

- Alimenter un SGBD à partir de documents structurés SGML produits par NEURODOC ou SDOC, constituant ainsi une base de données dite base infométrique car elle rassemble et organise des données bibliographiques normalisées et codifiées et les résultats de l'applications des différentes techniques d'analyse selon une structure de type relationnelle adaptée au calcul d'indicateurs quantitatifs et qualitatifs permettant d'évaluer et de comparer le positionnement thématique des acteurs de la recherche.

- Générer une interface WWW-SGBD pour l'analyse de l'information. Cette interface doit favoriser les interactions entre les schémas mentaux de l'usager et différentes représentations de l'information. Pour atteindre cet objectif, un hypertexte généré par ***HENOCH*** propose deux types de navigation complémentaires sur le Web :

- Une exploration intuitive basée sur l'utilisation d'une carte.
- Un mode de recherche orienté par la question "qui fait quoi, où, avec qui, quand, dans quelles sources (revue, congrès, ...)". Dans les deux cas, la navigation est assurée par l'exécution de requêtes SQL sur la base de données infométriques.

Le système HENOCH, de sa conception à son utilisation, est décrit en détail dans les chapitre 6 à 8.

3 Conclusion et articulation des chapitres suivants

Ce chapitre a défini la problématique de l'analyse de l'IST en la situant dans un contexte social : l'évaluation de la recherche et la veille scientifique. Il a montré en premier lieu que, **sur le plan opérationnel, l'analyse de l'IST s'appuyait sur différentes techniques (linguistiques, classificatoires, cartographiques) et des méthodes issues de l'infométrie, comme par exemple, la méthode des mots associés.**

Sur le plan informatique, cela s'est traduit par le développement d'une plate-forme logicielle, développement auquel j'ai largement participé (SDOC et HENOCH). Il reste que si **le processus d'analyse de l'information est un mélange d'exploration informelle intuitive et d'exploitation méthodique de l'information élaborée par différents outils d'analyse, il est nécessaire d'expliciter précisément comment peut s'effectuer cette exploitation pour pouvoir traduire cette démarche sur le plan technologique.** Les chapitres 2, 3 et 4 illustrent divers aspects techniques et méthodologiques d'une démarche générale d'analyse et d'interprétation des résultats qui s'est affinée progressivement dans le cadre d'études[15] menées dans différents domaines (sciences sociales, sociologie, physique). Dans les trois études décrites, la méthode infométrique utilisée est la méthode des mot associés. J'ai étudié cette méthode de manière approfondie, sur le plan de la démarche, sur le plan de son paramétrage (ce qui, sur le plan informatique, s'est traduit par l'outil SDOC) et sur le plan de l'exploitation de ses résultats. Le chapitre 2 met l'accent sur la nécessité et l'intérêt d'utiliser les statistiques bibliométriques en amont de cette méthode. Le chapitre 3 montre comment l'emploi de certaines techniques linguistiques permet d'améliorer et d'enrichir substantiellement les résultats obtenus par cette méthode. Le chapitre 4 montre plus particulièrement comment l'utilisation traditionnelle du diagramme stratégique dans la méthode des mots associés peut être complétée par une analyse des relations inter-thèmes sur une carte thématique en s'appuyant sur un hypertexte généré automatiquement selon une technologie antérieure au World Wide Web.

Le chapitre 5 constitue une articulation essentielle entre les trois premiers chapitres et les trois suivants. Il explicite la démarche d'analyse et de qualification des résultats applicable à deux méthodes de classification et cartographie de l'information qui sont décrites en détail : la méthode des mots associés, et une autre plus récente associant une technique de classification, les K-means axiales, à une technique d'analyse factorielle courante : l'Analyse en Composantes Principales (ACP). En mettant en évidence le besoin de pouvoir croiser dynamiquement certaines informations relatives aux résultats de classification et aux données à analyser, ce chapitre introduit en quelque sorte, les trois chapitres suivants, qui ont trait à la génération automatique d'hypertexte

[15] Études que j'ai effectuées ou auxquelles j'ai participé en collaboration avec des spécialistes du domaine.

dynamiques pouvant assister l'utilisateur dans sa démarche d'analyse de l'IST.

En conséquence de ce besoin, et cela a été signalé dans ce premier chapitre, un environnement d'analyse de l'IST devrait comporter, non seulement un ensemble d'outils d'analyse disponibles au sein d'une plate-forme, mais également un 'observatoire', véritable système d'information que nous appelons base infométrique, où l'utilisateur peut stocker, explorer et exploiter méthodiquement (par des requêtes) les résultats quantitatifs ou qualitatifs de l'application de différentes méthodes d'analyse sur des données brutes relatives à une problématique particulière.

Le chapitre 6 décrit et justifie une approche tout à fait originale (au moment de sa conception en 1995 [GRIVEL 1995b]) pour mettre en place un tel observatoire. Cette approche est basée sur une modélisation relationnelle des données et une architecture mixte : système de gestion de base de données et Web. Elle est opérationnelle au sein du système dénommé HENOCH.

Le chapitre7 montre comment HENOCH peut aider à construire des bases de données infométriques hybrides (multi-sources, multi types de données) exploitables pour le calcul d'indicateurs à des fins d'analyse de l'information scientifique et technique.

Le chapitre 8 décrit sur un exemple une démarche d'analyse de l'IST à partir d'un hypertexte généré par le système HENOCH. L'usager dispose de plusieurs modes de navigation conviviaux lui permettant de satisfaire de multiples besoins, comme par exemple, avoir une vue d'ensemble de l'organisation thématique d'un corpus de documents et de ses auteurs, identifier des relations inter-thèmes non explicites, identifier et regrouper les acteurs, les institutions, leurs vecteurs de communication (thèses, rapports, monographies, périodiques) par thèmes, évaluer le positionnement thématique d'un acteur, d'une institution, d'un pays, d'un mode de communication (périodique, congrès, ...), etc. Par un jeu de questions réponses, ce chapitre explicite le mode d'emploi de l'outil dans le cadre d'une étude sur les plantes transgéniques.

Pour conclure, le dernier chapitre 'Bilan critique et perspectives' permet, à partir d'une évaluation critique des fonctions du système par des utilisateurs, de dégager diverses voies de recherches possibles, notamment la visualisation et la comparaison dans le temps de représentations cognitives de données, la classification incrémentale, qui constituent de nouveaux enjeux pour la recherche sur la génération automatique d'hypertextes ergonomiques.

Bibliographie

[BALPE et al 1996] Balpe J.P, Lelu A., Saleh I. et Papy F. - Techniques avancées pour l'hypertexte - éditions Hermès, 1996.

[BOUTIN et al 1998] E. Boutin, B. Mannina, H. Rostaing, L. Quoniam Construction automatique de réseaux : un outil pour mieux appréhender l'information provenant d'Internet, Actes JADT 98, Coord. S. Mellet, UPRESA « Bases Corpus et Langages » Université de Nice 1998.

[BRADFORD 1934] Bradford S. C. 1934 - Sources of information on specific subjects - Enginering, 137 : 85-86, Janvier 1934.

[CALLON et al 1983] Callon M., Courtial J-P., Turner W.A., Bauin S. 1983 - "From Translation to Problematic Networks: An Introduction to Co-Word Analysis" in Social Science Information, vol. 22, pp. 191-235.

[CALLON et al 1986] M. Callon, J. Law and A. Rip (eds), Mapping the Dynamics of Science and Technology. London, Macmillan Press, 1986.

[CALLON 1993] Callon M. - La scientométrie - Que Sais-je, PUF Paris, 1993.

[CAPPONI 1999] Capponi Nicolas Généralisation de structures prédicatives. Application à l'analyse de l'information. Thèse de doctorat Science de l'information et de la communication, Université H. Poincarré Nancy 1, 1999.

[COURTIAL 1990] Courtial J.P. - "Introduction à la scientométrie : de la bibliométrie à la veille technologique", Anthropos - Economica, Paris.

[DKAKI et al 1997] Dkaki T., Dousset B., Mothe J. "Mining information in order to extract hidden and strategic information", Computer-Assisted Information Searching on Internet, RIAO97, pp 32-51, June 1997.

[DKAKI et al 1998] Dkaki T., Dousset B., Mothe J. "Analyse d'informations issues du Web avec Tétralogie", VSST'98 Veille Stratégique Scientifique & Technologique, Toulouse ,Octobre 1998.

[DOU 1995] Dou H. Veille technologique et compétitivité, Dunod, 1995.

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

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

[FRANÇOIS 1998] François C. - NEURODOC : un outil d'analyse de l'information -, Conférence. VSST'98 (Veille Stratégique Scientifique et Technologique), Toulouse, 19-23 octobre 1998.

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

[GODIN 1995] Godin R. Mineau G. Missaoui R. Mili H. Méthodes de classification conceptuelles basées sur les treillis de Gallois et applications, Revue d'intelligence artificielle Vol. 9, n°2, pages 105-137.

[GOLDFARB 1990] GOLDFARB C. The SGML Handbook, Oxford, Oxford University Press. 1990.

[GRIVEL et FRANCOIS 1995a] GRIVEL L., FRANÇOIS C. "Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique", *SOLARIS* n° 2, Presses Universitaires de Rennes, p. 81-112 (1995); et dans http//www.info.unicaen/bnum/jelec/Solaris.

[GRIVEL et FRANCOIS 1995b] GRIVEL L., FRANÇOIS C. Conception et développement d'un système d'information dédié à la veille scientifique, basé sur les sorties des outils de classification thématique : SDOC et NEURODOC , In : BALPE J.P, LELU A., SALEH I,.Eds, Hypertexte et hypermedia, réalisations, outils et méthodes, Paris, Editions Hermès: 109-118.

[GRIVEL et al. 1997] GRIVEL L., POLANCO X., KAPLAN A. 'A computer system for big scientometrics at the age of the World Wide Web', Scientometrics, vol.40, N°3 (1997), 493-506

[GRIVEL 1999] GRIVEL L. 'HENOCH, un outil d'analyse de corpus d'information scientifique et technique', Le Micro Bulletin Thématique n°3, L'information scientifique et technique et l'outil Internet, Editeur CNRS-DSI, 1999, p.27-44.

[HERWIJNEN 1990] HERWIJNEN E. "Practical SGML", Kluwer Academic Publishers, 1990

[HUOT 1992] Huot C. Analyse relationnelle pour la veille technologique : vers l'analyse automatique des bases de données, thèse de doctorat en Sciences de l'Information et Communication, Université Aix Marseille III, 1992.

ISO 8879 - 1986. Information processing - Text and office systems - Standard Generalised Markup Language (SGML), 155 pages

[JACQUEMIN 1994] Jacquemin, C. - FASTR: A Unification-based Front-end to Automatic Indexing - RIAO 94 Conference Proceedings «Intelligent Multimedia Information Retrieval Systems and Management», Rockfeller University, New York, October 11-13, p. 34-47.

[JACOBIAK 1996] Jacobiak F. L'information scientifique et technique, Que Sais-je, 1996.

[JACOBIAK 1992] JAKOBIAK. F. Exemples commentés de veille technologique. Paris : Les Editions d'Organisation, 1992, p. 27.

[KISTER et al 1993] KISTER J., RUAU O., QUONIAM L., DOU H. Application des outils bibliométriques en chimie analytique 4 ème Journées sur l'information élaborée Ile Rousse, Revue Française de bibliométrie 12, p. 437-456

[KOHONEN et al. 1995] Kohonen T. Kaski S. Lagus K. Honkela T. - Very large two level SOM for the browing of newsgroups - 5th International WWW Conference Paris 1995.

[KOPCSA et SCHIEBEL 1998] Kopcsa A. et Schiebel E. - Science and technology mapping : a new iteration model for representing relationships - Jasis 49 (1) :7-17 1998.

[KRUSKAL 1964] Kruskal J.B. - Multidimensonal scaling by optimizing goodness of fit to a nonmetric hypothesis - Psychometrika, 29 :1-28, 1964.

[LAFOUGE 1998] Mathématiques du document et de l'information. Bibliométrie distributionnelle, Habilitation à diriger des recherches, RECODOC, Univ. Lyon 1, Oct. 1998

[LEBART et SALEM 1988] Lebart L. Salem A. - Analyse statistique des données textuelles -, DUNOD, Paris 1988, 207 pages.

[LECOADIC 1994] Lecoadic Y. - La science de l'information - Que Sais-je, PUF Paris, 1994.

[LELU 1993] Lelu A. - "Modèles neuronaux pour l'analyse de données documentaires et textuelles" Thèse de doctorat de l'université de Paris VI. 4 mars 1993, 238 pages. -

[LELU et al 1997] Lelu A. , Tisseau-Pirot A.G., Adnani A.'Cartographie de corpus textuels évolutifs : un outi pour l'analyse et la navigation' Hypertextes et Hypermedia, Vol1. N°1, éditions Hermès, Paris, 1997

[LELU et al 1998] Lelu A., Halleb M., Delprat B. 'Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-Grammes » Actes des 4è journées internationales d'analyse statistique des données textuelles, Nice 1998.

[LEVEILLE 1998] Leveille V., Rostaing H., Quoniam L.Création d'hypertextes automatiques appliqués à la veille, VSST'98 Veille Stratégique Scientifique & Technologique, Toulouse ,Octobre 1998.

[LEVY 1990] Levy P. 'Les technologies de l'intelligence, Collection Points Sciences, Edition La découverte, 234p, 1990.

[LOTKA 1927] Lotka A.J. The frequency distribution of scientific productivity. Journal of the Washington academy of sciences, 16(2) :317-323, Juin, 1927.

[MARCOTORCHINO 1991] Marcotorchino F. , seriation problems : an overview, Applied stochastics Models and Data Analysis, Vol. 7 N°2, 1991),

MARTEAU P.F., KRUMEICH C. Analyse sémantique pour le veille technologique, IDT. Information, documentation, transfert des connaissances, Paris France; Pp. 258-263, 1995;

[MICHARD 1998] MICHARD A. 'XML Langage et application' Editions Eyrolles, 361 p, 1998

[MICHELET 1988] Michelet B. L'analyse des associations. Thèse de doctorat en Sciences de l'information, Université de Paris VII, 1988.

[MOED 1996] MOED H.F. 'Differences in the construction of SCI Based Bibliometric Indicators among Various Producer : A first Overview' , Scientometrics, , vol.35, N°2 (1996), 177-192

[NOYONS et VAN RAAN 1998] Noyons E., Van Raan A. Monitoring scientific developments from a dynamic perspective Jasis 49 (1) :68-81 1998.

[POLANCO 1995] Polanco X. 'Aux sources de la scientométrie', SOLARIS, Vol 2 «Les sciences de l'information : bibliométrie, scientométrie, infométrie, sous la direction de Jean-Max Noyer ». Edition : Presses Universitaires de Rennes, 1995, pp.13-78.

[POLANCO 1997] Polanco X. -La notion d'analyse de l'information dans le domaine de l'information scientifique et technique -, Colloque INRA, 21-23 octobre 1996, Tours. P. Volland-Neil, coord. *L'information scientifique et technique : Nouveaux enjeux documentaires et éditoriaux* ; Paris, INRA, 1997, pp. 165-172.

[POLANCO et al. 1997] POLANCO X., FRANÇOIS C., KEIM J.P. Artificial Neural Network Technology for the classification and Cartography of Scientific and Technical Information, to be published in Proceedings 6th International Conference of the International Society for Scientometrics and Informetrics, Jerusalem, June 16-19 1997.

[POLANCO et al. 1998] POLANCO X., FRANÇOIS C., OULD LOULY A. « For Visualization-Based Analysis Tools in Knowledge Discovery Process : A Multilayer Perceptron versus Principal Components Analysis - A Comparative Study », J.M. Zytkow and M. Quafafou (eds) *Principles of Data Mining and Knowledge Discovery*. Second European Symposium, PKDD'98, Nantes, France, 23-26 September 1998. Lecture Note in Artificial Intelligence 1510. Subseries of Lecture Notes in Computer Science. Berlin, Springer, pp. 28-37, 1998.

[QUONIAM L. 1988] Quoniam L. Bibliométrie Informatisée et Information Stratégique, Thèse de doctorat. en Sciences de l'information et de la communication. Université Aix-Marseille III.. pp. 330, 1988.

[QUONIAM L. 1992] Quoniam L. Bibliométrie sur références bibliographiques: méthodologie in: La Veille Technologique: l'Information scientifique, technique, industrielle. DUNOD, 1992.

[Rapport Inria N° 3198] - Acquisition et structuration des connaissances en corpus : éléments méthodologiques - Muller C., Polanco X ., Royauté J. Toussaint Y. Rapport Inria N° 3198.

[ROSTAING 1996] ROSTAING H. 'La bibliométrie et ses techniques', Edition : sciences de la sociéte, coll : « Outils et méthodes », 1996, 131p.

[ROUSSEAU 1998] Rousseau F. - L'analyse de corpus d'information comme support de la veille stratégique - Document numérique (2), 177-202, juin 1998 .

[ROYAUTE et JACQUEMIN 1993] Royauté, J. et C. Jacquemin (1993), "Indexation automatique et recherche de noms composés sous leurs différentes variations". Informatique & Langue Naturelle, ILN'93, Nantes, France.

[ROYAUTE 1999] ROYAUTE J. Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information, Thèse de doctorat en informatique, Université H. Poincaré Nancy I, 1999.

[SERRES 95] Serres A. L'hypertexte, une histoire à revisiter, Documentaliste, 1995, vol 32 n°2, 71-83.

[SMALL 1997] Small H. - Update on science mapping : creating large document spaces Scientometrics - 38 (2) : 275-293, 1997.

[SMALL 1999] Small H. - Visualizing science by citation mapping- Jasis 50 (9) :799-813, 1999.

[WOLFRAM 1996] Wolfram D. Inter-Record linkage structure in a hypertext bibliographic retrieval system Jasis 46 (10) :765-774, 1996.

[ZIPF 1949] Zipf G.K. - Human Behavior and the Principle of Least Effort - Addison-Wesley, 1949.

[ZITT 1996] ZITT M. , TEIXEIRA N. ‘Science Macro-Indicators : some aspects of OST Experience Scientometrics’, vol.35, N°2 (1996), 209-222.

Chapitre 2[1]

Bibliométrie et cartographie de l'IST par la méthode des mots associés : démarche applicative

L'analyse de l'information peut être définie comme l'application de techniques de traitement automatique du langage naturel, de classification automatique et de représentation graphique (cartographie) du contenu cognitif et factuel des données bibliographiques.

Même ainsi outillée, l'analyse de l'IST ne peut être effectuée sans s'appuyer sur une solide démarche méthodologique. Ceci suppose une documentation adéquate de la méthode employée et de la chaîne de traitement, une définition claire des sources de données et des indicateurs utilisés. C'est l'approche qui est suivie dans ce chapitre pour illustrer l'utilisation des lois bibliométriques pour l'analyse de l'information par la méthode des mots associés.

La loi de Bradford est appliquée pour définir les fichiers de données qui seront en entrée du processus de classification et cartographie dans le cadre d'une application dans le domaine des sciences sociales. La méthode des mots associés est employée pour structurer l'information en thèmes et représenter ces thèmes et leurs relations dans un espace bi-dimensionnel.

Une documentation de la méthode est proposée : principes, paramétrage, variables utilisées pour décrire les caractéristiques des thèmes et les représenter géographiquement sur une carte.

Les résultats obtenus sont discutés et notamment la perspective de construire des cartes capables de représenter et visualiser l'état de la connaissance scientifique à partir des bases de données. La cartographie de la science est en effet une représentation spatiale de la manière dont les disciplines, les domaines, les spécialités, les articles, les auteurs sont associés les uns aux autres. Un peu à la manière dont des cartes géographiques peuvent rendre compte des relations entre des caractéristiques physiques ou politiques.

---

[1] Polanco X., Grivel L. -'Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United State of America)', JISSI International Journal of Scientometrics and Informetrics, Vol.1, Nr. 2, june 1995, pp 123-137. 4th International conference of Bibliometrics, Informetrics and Scientometrics -, Berlin, Germany. 1993

1. INTRODUCTION

We group bibliometrics as well as scientometrics under informetrics. What is Informetrics for? In our field of performance, informetrics operates the following functions: analyzing, assessing and mapping scientific and technical information (STI). The analysis is aimed at answering strategic needs and serving scientific and technical monitoring purposes. The end product is “information on information”. There are two kinds of STI assessment: a metrical assessment of information flows (articles, journals, reports, patents), and a qualitative assessment of the information processed (relevance). Mapping (or graphical representation) consists in presenting STI as maps on which to position both information contents and research actors.

Moreover, informetrics is for us a research programme in the context of an information industry. The Institut de l’Information Scientifique et Technique (INIST) is an integrated information centre, created by the French Centre National de la Recherche Scientifique (CNRS) for worldwide promotion of French and European research. Its mission is to collect and process the results of research and to make them immediately accessible.

Scientometric analysis has mostly been applied in the natural and life sciences. A small number of studies have used scientometric tools to analyse the research developments in the social sciences. Whereas scientometric tools have proved their usefulness as monitors of research developments in the natural and life sciences, evidence on this point is lacking almost completely for the humanities and social sciences disciplines. This paper is an attempt to apply a scientometric approach in the field of the social sciences, and to evaluate its potential usefulness.

The first goal of the study is to map knowledge or "subject maps" as Price said (1986, p.269). According to Small and Garfield (1988, p.46): "The notion that science can be mapped was first clearly stated by D. Price during the 1960s". In order to map knowledge, we use co-word analysis (Callon, Law, Rip 1986). We have implemented (SDOC programmes) the co-word analysis in order to classify and visualize the STI. It is based on the keywords assigned to scientific documents. As a general definition, we shall take a co-word map of scientific information to be the representation of the topology of relationships between distinct subject areas or research themes, which are embedded in the database from which the data has been extracted.

In this paper, we are going to describe the application of our informetric chain (based upon the analysis and processing of word associations in a database) to the social sciences information, in the specific field of sociology. For this purpose, we use the FRANCIS database produced by INIST in France. FRANCIS is a unique set of 20 multidisciplinary bibliographic data bases covering the core of the world literature in Humanities, Social Sciences and Economics. Then, we shall limit our analysis to sociology information just as it is stored in a particular database. We will focus our attention on the results of the treatment of the four sets of bibliographic data, each corresponding to one of the following publishing countries: France, Germany, United Kingdom and United States of America. We emphasize that this four-country comparison does not represent a complete survey of the state of the art

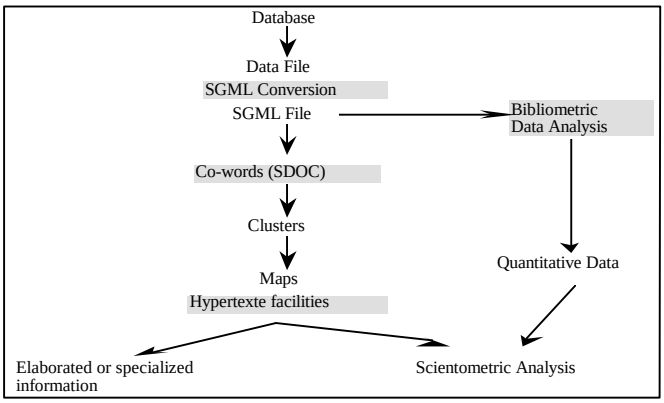


Figure 1: The informetric chain operates at a number of levels

2. METHOD.

2.1. Co-words analysis.

Co-words analysis is an alternative tradition to a more well known and wide-spread tradition of citation and co-citations analysis in the scientometric community. The idea to use keywords to describe the distribution of units of information in a scientometrics sense, is due to the Centre de Sociologie de l'Innovation de l'Ecole de Mines de Paris. The first article in a journal describing this method was published in 1983 (Callon and alii, 1983; 1986).
According to indexing documentary tradition, a keyword is an indicator of the subject content of a document. We are ready to admit that the problem here is the quality of the indexing process. This problem is known as the “indexer effect” (see Healey and alii, 1986, p. 245; see also Polanco, 1993). It is important to note what the main characteristics of the method are. As we know, the first quantitative pattern of keywords is the frequency. Bibliometricians know that the frequency distribution of words is governed by Zipf law. The second pattern is the keywords co-occurrence. The idea of co-occurrence is central. Co-words are, as its name indicates, a relationship indicator (as well as citation and co-citation); this cannot be expressed by an isolated word, as for instance the keywords of frequency one. The third level is the formation of clusters. A cluster is a group of associated keywords indexing a number of papers which are the units of information about the subject indicated by the cluster. In this sense, each cluster is an indicator of knowledge founded on frequency and co-occurrence statistical values. The last procedure is to map knowledge. Usually scientific knowledge is in the form of collections of bibliographic data. A representation is a set of conventions about how to describe information. Thinking seriously about scientific information of any sort requires thinking seriously about what representation is best suited to the domain from which the bibliographic information comes from. Indeed, the main purpose of the paper is to demonstrate the central importance of finding good representations capable of bearing good descriptions.

2.2. SDOC programmes.

SDOC is an original implementation of co-word analysis from the point of view of computer information technology. The programmes are implemented under the Unix

operating system and written in C language, according to a toolkit philosophy based on modularity and data flow communication rather than a “press-button software package”. Thus, the modules are organized in a toolbox (Ducloy and alli, 1991).
We use SGML (Standard Generalized Markup Language) to describe bibliographical references whatever their source; and SGML standard is used as pivot format and as specification language for coding intermediate data (see Figure 1).
Co-words analysis has been broken down into the following steps, each corresponding to an independent module communicating with others by file: (1) Building keywords index from a set of bibliographic references; (2) Computing cooccurrences of keywords and measuring the similarity of the keywords; (3) Cutting up the keywords associations network into clusters; (4) Classifying references into clusters; (5) Building scatter diagrams or maps. In the process, SGML is the specification language of data transmitted from one step to another. The modules are parameterized by the input and the output filename and its specific parameters. Statistics tools and visualization tools are available to assess the results. Finally, the hypertext technique provides the analysts with an interactive working tool (Grivel and Lamirel, 1993).
The clustering algorithm which groups the associated keywords into clusters is an adaptation of the single link clustering algorithm. After the clustering process, the documents are assigned to the clusters.
In order to permit an easier consultation and interpretation of the clustering and mapping results, SDOC converts the “knowledge network” represented by the clusters into hypertext nodes and links. The goal of this conversion is to allow the user to visualize very quickly the knowledge organization of a topic, the key figures, the organisations.

3. DATA & BIBLIOMETRIC ANALYSIS.

We use the bibliometric analysis in order to build the datafile that will be the input of the automatic clustering and mapping process (SDOC programmes). This bibliometric analysis is based on some bibliometric indicators, as for instance, the document type, the date of publication, and the country of publication.

3.1 Construction of the data file.

A funnel-shaped step by step process is applied on the basis of the selected bibliometric indicators as criteria of construction of the final “target datafile”. The main steps of this funnel-shaped process are: (1) the extraction of a datafile from the FRANCIS database; (2) from this source datafile a first raw datafile is constructed by means of the document type bibliometric indicator application; (3) a second datafile is extracted from the former by the application of the country of publication as criteria of selection (4) and finally, the input datafile to SDOC programmes results from the application of Bradford’s law.
The first step in a informetric analysis is to extract the “target literature” from an information retrieval database (Turner and alii, 1988). In the present case, we simply start from the literature collected and processed from 1989 to 1991 by the FRANCIS database on sociology. The size of this datafile is of 13.942 records; there are different types of documents: journal articles, books, reports, proceedings and Ph.D. This is our starting raw datafile from which we shall build a second datafile. The decision here was to focus on the journal articles.
There are predominantly journal articles in our source datafile, they represent 84% of the sociological information stored in the FRANCIS database from 1989 to 1991. There are

720 journals from which 11 661 articles originate. The other sources are books, reports, proceedings, and theses, at 16% of the raw datafile. The reports and theses essentially concern French sociology.
The date of publication of this literature corresponds mainly to the last years of the 1980's sociology. (13.735 records , 98%, between 1986 and 1991).
The authors' institutional affiliation does not appear in the FRANCIS references, so we have used the publishing country of journals for the definition of our "target" literature. As we can see in Table 1., the journals of the four publishing countries selected and the articles issued from these sources represent 70% of the total. The remaining 30% is distributed over 44 publishing countries in the world. This is a long-tailed distribution, and as we know, this type of distribution appears to be characteristic of bibliometrics. We decided to focus our analysis on this literature of the four publishing countries . In this set, France is overrepresented, Germany and United Kingdom are approximately equal, and the United States are well represented.

Table 1.

Publishing country	Number of journals	%	Number of references	%
France	270	37.55	3245	27.83
Germany	39	5.42	860	7.38
United-Kingdom	49	6.82	1310	11.23
United-States	143	19.89	2787	23.90
Total	501	69.68	8202	70.33
All publishing Countries: 48	719	100.00	11661	100.00

Considering the excessive impact of two journals in the set of 501 journals selected, and their specialized nature, we decided to treat them separately and to remove them from the "target literature". They are *Economie et Statistique* (France), and *Journal of Marriage and the Family* (Etats-Unis). The clustering of the 249 articles of *Journal of Marriage and the Family* (Etats-Unis) provided 21 clusters and we obtained 11 clusters from the 138 articles of the journal *Economie et Statistique* (France). This case underlines that, if the number of references is statistically significant, one can proceed to a content analysis of these references using SDOC, as the one we discuss in section 4 of this paper. We shall not present here the results obtained from these two journals.
So our "target literature" becomes at last : France 269 journals as a source of 3.107 references; United-States 142 journals as a source of 2.538 references. Germany and United-Kingdom remains unchanged.

3.2. Application of the Bradford Law.

To select the "target literature", we applied "Bradford's law of rank distribution". The four-country sociological journals were ranked by decreasing productivity of articles. Then for each country, we defined 4 groups (see Table 2) composed of the most productive journals so that their cumulative number of references reach respectively 25%, 50%, 75% and 100% of the corpus.

Table 2.

Publishing country	Number of references	Number of journals with % of references			
		25%	50%	75%	100%

France	3107		12	31	68	269	
Germany		860		3	6	15	39
United-Kingdom	1310		3	8	15	49	
United-States	2538		7	17	32	142	

We defined as “nuclear zone” the journals which produce 50% of the references. We focused on the four-country nuclear zone (Table 3), in order to map the most important publications, of course according to FRANCIS database coverage on sociology.

Table 3.

	(S)		(R)	
Publishing Country	Journals	%	References	%
France	31	11	1568	50
Germany	7	15	462	53
United Kingdom	8	16	676	52
United States	17	12	1287	51

So, the “target” bibliographical data, that we used for the mapping process, is not only a set of sources (S) and a set of references (R), but also the application of a function expressing the source-reference relationships; it is the Bradford’s ranking analysis. From the point of view of the date of publication, the nuclear zone is a sociological literature published during 1987-1990.

France is over-represented comparatively to other publishing countries. It may be an expression of the wish of exhaustivity to cover national literature. A certain eclecticism is expressed by a two-level literature: one is more strictly scientific or academic, the other one corresponds more to an enlightenment literature. The category of enlightenment publications includes popularizing articles and reviews in magazines.We take the distinction between ‘scholarly’ and ‘enlightenment’ publications from Nederhof and alii, (1989, p. 427-428). This is not the case for the other countries where the journals selection appears much stricter. These facts only express a policy of coverage of journals. We cannot use these data to compare countries' productivity. The inequality existing in the productivity of the four countries is not a problem for the goals we have fixed in our introduction.

4.RESULTS AND COMMENTARY

The obtained results are presented in two parts. The first one is dedicated to the presentation of the lists of clusters and the second one to the mapping of the clusters on scatter diagrams. It corresponds to two phases of the method. In the first phase, it is a question of structuring information and identifying the emerging research subjects (cluster analysis). The second phase is the graphic representation of these subjects in a two-dimensional space (network analysis).

SDOC	Analytical Action	Object Study
Automatic Classification → Clusters	Cluster Analysis	Research Subjects or Themes
Graphic Representation on two-dimensional space (y,x) → Maps	Network Analysis	Global & Local Networks

Figure 2: Human-Machine Information Processing.

Figure 2 allows us to distinguish two other phases concerning the information processing, (1) a first machine-based phase, the SDOC application, and (2) the phase where there is the action of an expert or knowiledgeable person. Our information processing is based on cluster and network analysis techniques, in consequence the expert's goal is to study the themes and networks. In this second phase, hypertext represents an analytical tool which allows navigation through the information space of clusters and networks.

4.1. Cluster analysis

Cluster analysis is, as we know, the generic name for a wide variety of procedures that can be used to create a classification. The procedure empirically forms clusters or groups of key words. The clustering method is a multivariate statistical procedure that starts with a bibliographical data set containing information about a subject and attempts to reorganize the bibliographical information into relatively homogeneous groups. As we have already noted in section 2, the coword clustering method (implemented by SDOC programmes) is designed to create groups or clusters of associated keywords (co-words) as a means to indicate some numbers of research themes. In this particular application on sociology data file, we have applied the Equivalence Index. If we call Cij the cooccurrence number of two keywords *i* and *j*, Ci and Cj their occurrence numbers, the Equivalence Index (Eij) is given by the following equation:

$$Eij = Cij^2 / (Ci \times Cj).$$

The clustering algorithm which groups the associated keywords into clusters is an adaptation of the single link clustering algorithm. All the elements which are to be initially clustered constitute a large flat association network, i.e. a system of relationships where the keywords are related to each other. The separation of the association network into clusters is done according to a readability criteria: the cluster size (minimum and maximum number of components) and the number of associations in the cluster. If a pair of terms belongs to the same cluster, the association between the terms is an internal association. If they belong to two different clusters, the algorithm tries to aggregate the clusters into one by merging them. The merger is authorized if the size of the resulting new cluster respects the "readability criteria". If not, the association is considered as an external association. In this application, the parameters for each datafile were : minimal size of the clusters = 4 keywords; maximal size of the clusters = 10 keywords; maximal number of external associations = 10; maximal total number of associations = 20.

After the clustering process, the documents are associated to the clusters. A document is related to a cluster if, within its indexing terms, there is at least one pair of terms which can constitute either an internal association or an external association. We associate a list of authors, and a list of document sources to each cluster, as this information is available in the studied datafile.

Number of lines	Definition of the statistical parameters
[1]	Minimal cooccurrence of keywords (cooccurrence threshold)
[2]	Initial number of documents
[3]	Number of documents with at least a couple of keywords satisfying [1]
[4]	Number of clusters
[5]	Number of documents in the clusters

[6] Number of documents appearing only in one cluster

Table 4.

	France	Germany	United Kingdom	United States
[1]	4	2	3	4
[2]	1568	462	676	1287
[3]	1119	392	498	938
[4]	28	24	17	20
[5]	944	324	434	756
[6]	493	156	233	422

These are the main global indicators which allow us to adjust the clustering process by measuring the loss of information in function of the cooccurrence threshold and then the ratio number of references in the clusters / initial number of references. Table 4 provides only the data corresponding to our final choice for that application. We have tried to find a good compromise between the number of clusters for each data file and the loss of information due to both the selected cooccurrence threshold and the clustering parameters.
The statistical variables which characterize each cluster are the following:

Number of columns	Definition of the statistical parameters
[1]	Cluster's saturation threshold
[2]	Density, the mean of the internal associations
[3]	Centrality, the mean of the external associations
[4]	Number of keywords defining the subject
[5]	Number of internal associations (between the keywords defining the subject)
[6]	Number of external associations with other subjects (or clusters)
[7]	Number of citations of a subject by other subjects
[8]	Subject's bibliographic information (number of references)
[9]	Specific subject's bibliographic information

We indicate for each cluster the quantitative value of these parameters. The values of the first three columns [1] [2] [3] in the tables below are obtained by the Equivalence index ; those of the columns [4], [5], [6] are the size parameters of clusters which results from parameters fixed a priori for building clusters. The values of the last two columns [8] [9] concern documents classification by clusters. These are the indicators which allow us to characterize the clusters.
In the tables 5 to 8 in the appendix, each cluster is a row and each statistical parameter a column. Then we can choose a parameter, and rank the clusters according to their quantitative values in the selected column. Here, the clusters have been sorted by [2] *density value*, the mean of the internal associations which characterizes the strength of the links between the words making up the cluster (intra-cluster associations). The stronger these associations are, the more the subject corresponding to the cluster constitutes a integrated unit of information (or knowledge). *Centrality* [3] measures, for a given cluster, the intensity of its external associations with other clusters (inter-clusters associations). The more of these associations there are, and the stronger there are, the more this cluster designates a subject that is considered important in the knowledge network. The word *citation* [7] is used to indicate the fact that one cluster has been

cited in the external associations of another cluster; When one cluster, by its external associations, refers to another cluster, the latter has been cited by the former as a related item of information. The bibliographic information represented by a given cluster is measured and characterized by the parameters [8] and [9]. The column [9] is also an indicator of the bibliographic independence of a cluster in relation to other clusters.
The name of a cluster is only a label. The heuristic used to label the clusters is to choose the keyword which appears the most frequently in the associations. The name of a cluster suggested automatically may sometimes be more a mask than a source of information. The program should allow an expert to change its name in this case, as,for instance, for the *Relations* cluster in the four lists of clusters. But taking into account that this cluster is related to a significant number of bibliographic references, SDOC programmes permit us to come back to this number of references, to isolate it in a datafile and to process this datafile in order to again obtain a classification of the information masked by the label, a visualisation of information. We call this action the *"russian doll" procedure*.
One can also see the use of the word *region* in the singular and pluriel forms. This demonstrates a certain indexing policy and indicates for us the need to adopt methods to normalize the indexing vocabulary in input in order to correct these undesirable effects.
These tables of clusters enable us to know something about the problems studied and their relative importance in the datafiles. We can then analyse in more detail each element, that is to say (1) the keywords which form one cluster, (2) the internal and external associations with other clusters, (3) the sources, (4) the authors and (5) the titles of articles belonging to clusters**.** The conversion of all this data into hypertext hugely facilitates these operations. It increases the analysis and assessment task performance of this information, previously structured by the automatic clustering process.
We can also compare the research subjects in each case; for instance to compare the European publishing research in sociology , to compare it as a whole with the United States, from the point of view of "study subjects" (similarities, differences) and areas of research as for instance social, economic or politics areas. We can also focus on a subject in the four countries (transverse analysis), for instance technological innovation or social deviances (see maps below).
Another possibility is to use the co-word clustering process as an instrument for bibliographic retrieval. Retrieval systems are designed to enable a user to query a database of documents or document surrogates. In this sense, we have a co-word based retrieval system, where the user can navigate through clusters in different subject areas of research and immediately identify their authors, journals, titles of papers.
Looking at the scatter diagrams is the next step of the co-word analysis. The scatter diagram for any set of keyword-clusters shows what we call a "knowledge space" (Meincke and Atherton, 1976), or "information space" (Brookes, 1980). In this space, clusters are the indicators of items of knowledge and their positions are indicators of the *density* (Y axis) and *centrality* (X axis) of this item of knowledge. Such diagrams are included in the next section of this paper. Each scatter diagram is a representation of a set of clusters using the values of the columns [2] and [3] of the tables in the appendix.

4.2. Representing Knowledge in Scatter Diagrams

From a perspective of analysis, the first stage of description was the cluster analysis, and now the second step is the network analysis. Relations are principally the subject of network analysis. A network is a type of relation linking a defined set of clusters (unit of

information). The clusters can be defined as micro-networks or graphs and the maps as macro-networks. They are the building blocks of our network analysis.
We propose a two dimensional device for visualizing the organization of objective knowledge diffused by bibliographic data (information). We develop a representation of information items. The chief ouput is a spatial representation, consisting of a configuration of subjects (or clusters), as on a map. Each subject in the configuration corresponds to one item of information. This configuration reflects the “hidden structure” in the data, and often makes the data much easier to comprehend.
Before going into details about the description, a remark must be made about the sense of the scatter diagrams in our procedure. We use them as a way to produce a knowledge representation. "A representation has been defined to be a set of conventions for describing things. Experience has shown that designing a good representation is often the key to turn hard problems into simplest ones, and it is therefore reasonable to work hard on establishing what symbols a representation is to use and how those symbols are to be arranged to produce descriptions of particular things" (Winston, 1977, p. 179).
On the other hand, as Poppers says (1979, p.108-109) there are two different senses of knowledge, the first is “knowledge in the subjective sense, consisting of a state of mind”, and the second is “knowledge in an objective sense, consisting of problems, theories, and arguments as such. Knowledge in this objective sense is totally independent of anybody's claim to know; it is also independent of anybody's belief, or disposition to assent; or to assert, or to act. Knowledge in the objective sense is knowledge without knower; it is knowledge without a knowing subject". Knowledge is taken by us in an objective sense, consisting of journal literature, the medium through which natural or social scientists report their own original work and in which they evaluate work done by others.
Two main categories of problems arise from the study of scientific knowledge. One deals with the act of producing knowledge; the other is concerned with the very structures of knowledge produced by scientific activity. (see Popper, 1979, p. 112-113). We are concerned in our study by this second category of problems. Co-word analysis is a way of mapping the structure of scientific knowledge expressed by authors in their publications.
What do maps actually represent? On the one hand, they represent a set of clusters which designate specific centres of interest or themes or subject areas. On the other hand, they represent a network structure. It is a two-dimensional space. The Y axis called "density indicator" is defined by the strength of the internal word associations. It is thought to indicate internal coherence of the subject area. The X axis called "centrality indicator" is defined by the strength of the external associations. It indicates the role of a subject area in structuring a field of research.
When Derek de Solla Price said that the pattern of bibliographic references indicates the nature of the research front, he was clearly thinking of the citation analysis (Price, 1965). The citation of one paper by another in its footnotes or bibliography was the basis of his idea that science can be mapped. The co-word analysis is another tradition in mapping science. We emphasize that co-word maps are representations of knowledge structures network .
The figure 3 shows that with two theoretically important attributes, *density* and *centrality*, we have four possible combinations (see Callon et alii, 1991, p.165-167).

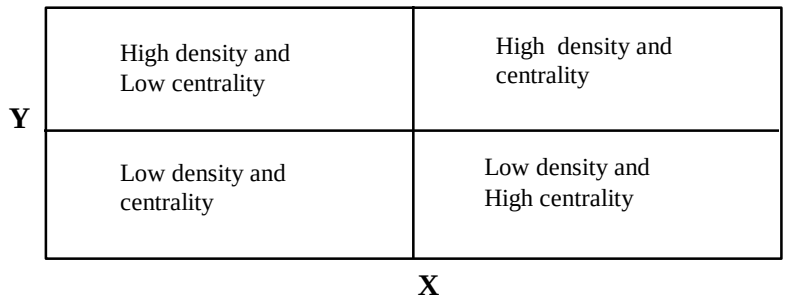
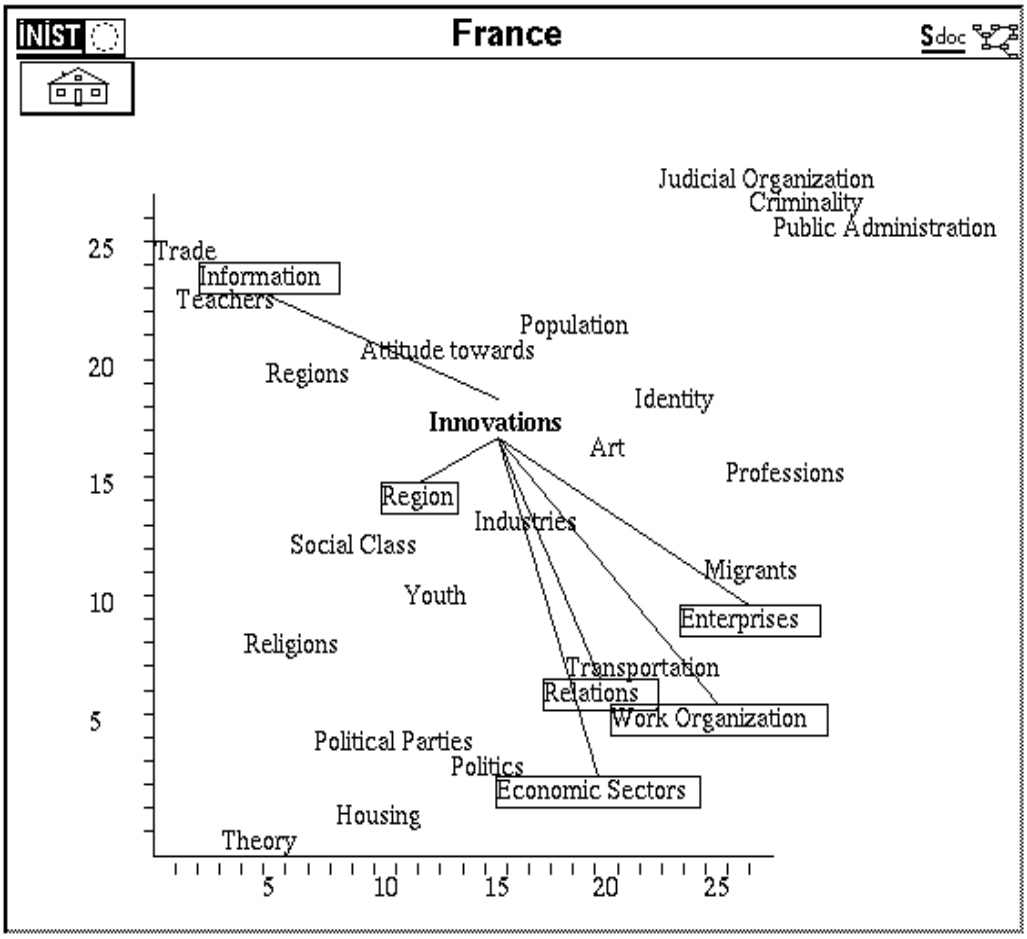


Figure 3: Scatter Diagram and categorized classes of clusters as indicators of research subjects identified from bibliographic data by computer programmes.

In examining a scatter diagram, the first thing to look for is the clusters distribution in these four zones of the diagram. The clusters are scattered according to the mean value of the internal associations (along the Y-axis), and of the external associations (along the X-axis). The information provided by the diagrams concerns the relative importance of themes or subjects (clusters) according to these two attributes: *density* and *centrality.* This relative importance of clusters is set up from the network of internal associations of each cluster (position along the Y-axis), and external associations between the clusters (position along the X-axis). The first value (along the Y-axis) defines categories of subjects more or less coherent and integrated as units of information. The second value (along the X-axis) defines more or less isolated or linked clusters, this is the notion of *centralness* of a theme in the knowledge space.

Our scatter diagrams are not metric spaces; the fact that two or three clusters are close to one another does not mean that they are closely linked to each other. On the other hand, we arrange the clusters by rank on the Y and X-axis. The number of ranks is equal to the number of clusters. So, the maps can be interpretated as rows on the Y-axis and columns on the X-axis.

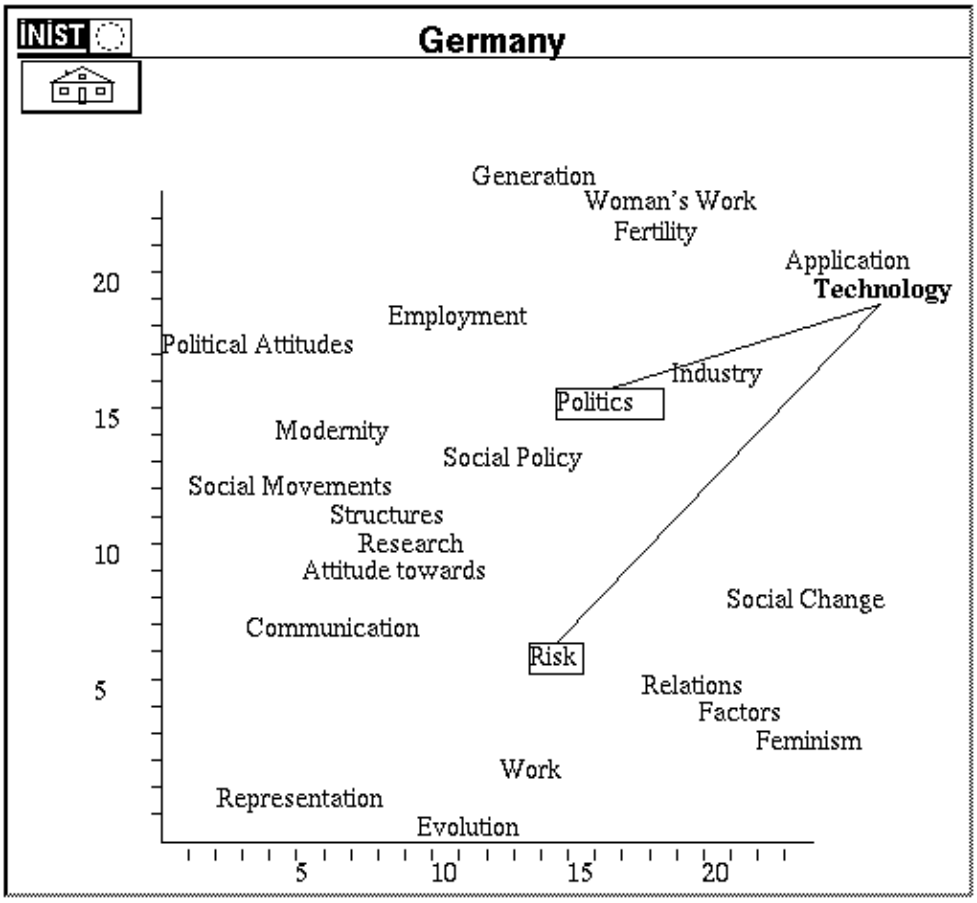
Now we are going to show how the maps can be used to help the analysis of research themes such as those linked with other themes forming a network in this way. Certainly, the analysis may descend to the level of the authors, sources and articles each time.



Map 1

The map 1 shows a set of three clusters together in a position of high density and high centrality. They are the themes about *Public Administration*, *Criminality* and *Judicial Organization*. This centrality is specially explained because they are closely connected, but at the same time each one represents an integrated internal unit of information on this subject (or high density). In reality, they represent an information area that is the result of the weight of certain specialized journals in security, criminology and laws in the sources of the data file. This area is open to *Politics* and *Professions* by means of the external associations of the *Judicial Organization* cluster.

The map 4, which gives a representation of the sociological literature published in journals edited in the United States, also highlights a dimension of social deviance. This is again the same phenomenon, that is to say the important weight of the sources of information specialized in these subjects.

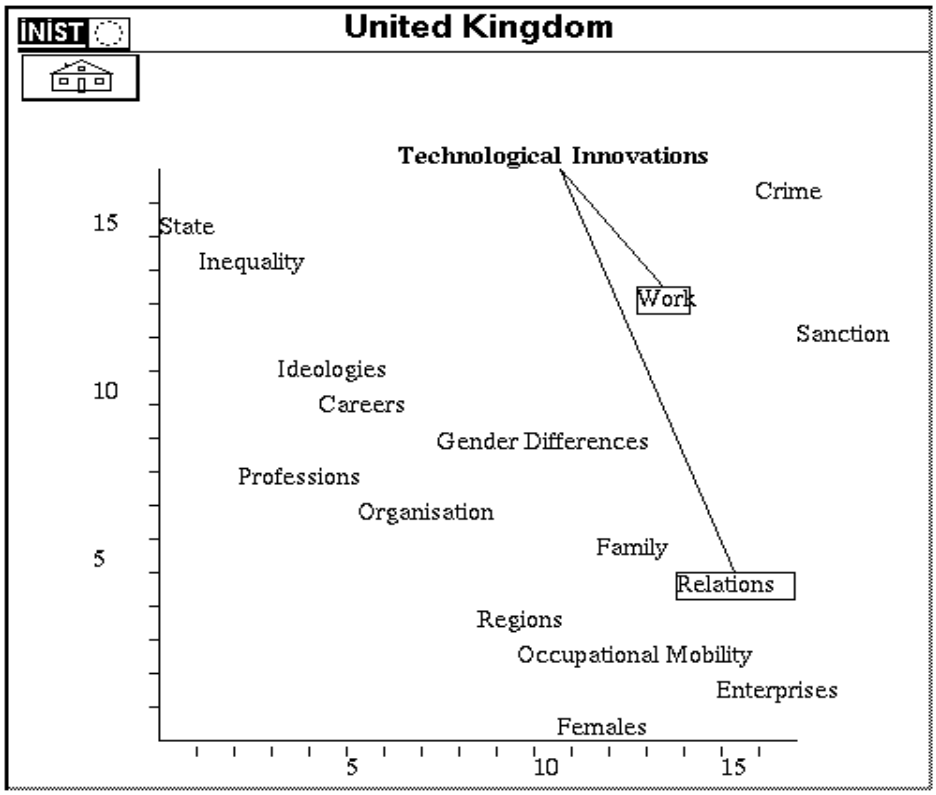


map 2

In the case of the European maps (maps 1, 2 and 3), our choice has been to show here the maps that highlight the associations of clusters as *Technology* or *Innovation* or *Technological Innovation* with the other clusters. This is in order to show how a cluster, that we consider as a graph, or in other words a micro-network, can be situated inside a larger network, macro-network or context. We can compare the position of the subject *Technology & Innovation or Technological innovation* in the three contexts.
In the case of France, this subject appears at an average position along the two axes, and is linked (new communication technologies) with *Information*, a subject whose position on the map indicates that it is cohesive but without centralness. At the same time, the other associations refer to significative clusters from the centrality point of view, but lowly integrated as unit of information. All these clusters constitute an economic area of sociological research. (There we also have the example of "indexer effect": a same concept is expressed in the singular and the pluriel form, *Region* and *Regions*, whenever they constitute a single and same concept).

On the map 2 concerning the German journals of sociology, the cluster *Technology (Innovation or Technological innovation)* is plot at a high value along the two centrality and density axes. By means of its internal associations, it exhibits three sectors (1) computerisation, (2) enterprise and industrial enterprise, and (3) human genetic engineering; the external associations refer to clusters *Politics* and *Risk* (more specifically the nuclear risk). In the cluster *Politics*, we have a junction concerning "mass media" and "public opinion".

On the France map, the technological innovation theme is linked to economic development and work organisation changes (also visible on the United Kingdom map). Whereas on the map 2 (Germany), this theme is associated to the risks and social impacts of the computerisation and the genetic technologies applied to human reproduction,.
Now, if we look at the map 3 United Kingdom map, the *Technological Innovation (or Innovation)* is a high density and high centrality cluster, associated with *Work* and *Relations.* Again, we find the ambiguous word *Relation*s as a descriptor and then as a label of a cluster. But the *"russian doll procedure"* is handy to visualize what is hidden under this subject because of the number of records aggregated in this cluster (128 records). *Work* is a cluster in which we find sociological studies on skill and deskilling problems because of the technological changes, and the *Work* cluster is associated by its external associations to the theme *gender differences*. This is the context in which the social studies of technological innovation are situated in our information space.



Map 3

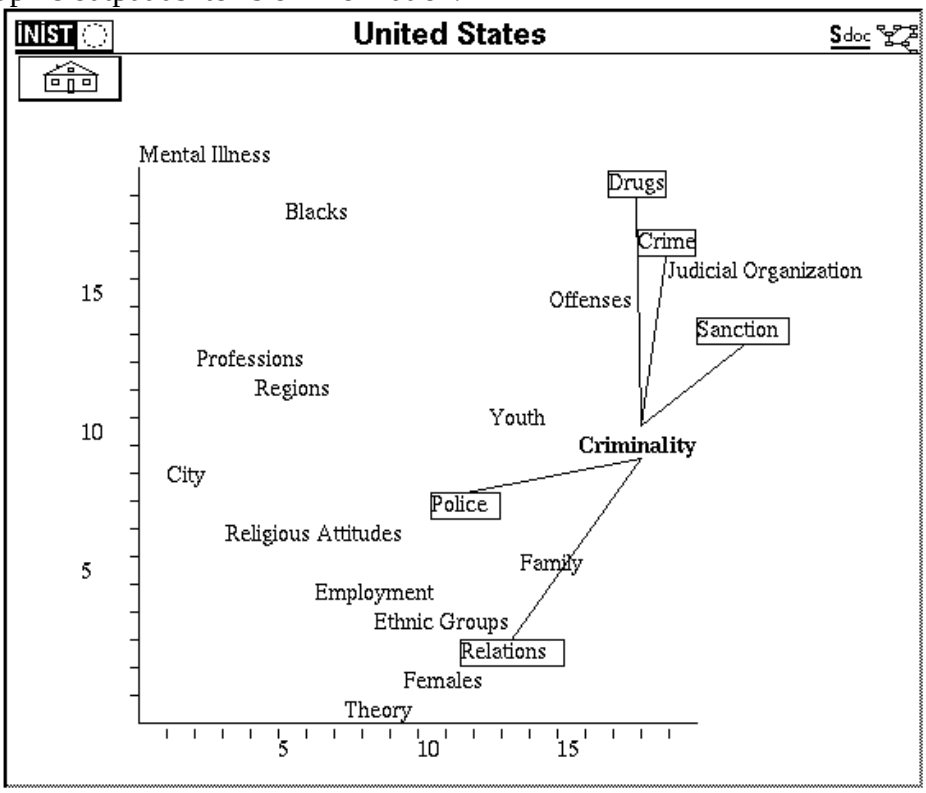
As in the case of the literature published in journals edited in France and the United-States, the United Kingdom map shows that the subjects *Sanction* and *Crime* stand out. This is an indicator of the relative importance of the sociological research dedicated to social deviance problems.
The United States map is a representation of the important weight of the specialized publications in social deviance and anomy. The network is a graphical representation of the information essentially published by the journals *Criminology, Crime and Delinquency* followed by *Social Forces* and *Social Problems*.
The information on technological innovation is not visualized on the map, this information is inside the *Regions* cluster, because the studies concern the agriculture, and their source is the *Rural Sociology* journal. On the contrary, the sociological studies on technological innovation published by European journals appear in an industrial context

of change and innovation (France, Germany, United Kingdom), also in the context of communication technologies (France, German) and human genetic engineering (German).

This brief overview of four cases of network analysis underlines the important emergence of the structural properties of the information. In addition, we can note the problems induced by the extreme sensibility of the co-word analysis to the quality of indexing. Beforehand an important effort of normalization is needed. A second problem is always to consider explicitly the type of journals which are at the origin of the information that we analyse and represent, especially when it is a question of representing the results of a given field of research.
Finally, it is interesting to underline that maps allow a user to visualize the knowledge structure of the document data file. The idea is to present information within a cognitive structure so that the experts can assess its validity. On the other hand, as Brookes says (1981, p. 10) : "As a map grows it will reach a stage at which it could be used as a database".

5. CONCLUSION.

We would like to stress two main purposes concerning our approach. The first is to map knowledge structures, and the second is to watch science activity by means of its bibliographic output as items of information.



Map 4

Mapping knowledge structures : this discussion about knowledge and information spaces provides a perspective, the production of cognitive maps of any developing knowledge field stored in the database at any time. Furthermore, SDOC programmes rely on the hypertext paradigm to represent the thematic maps, and allow the user to navigate

through a hyperspace composed of clusters, relationships between clusters, documents related to these clusters, and so on. Such a hypertext map would become of strategic interest to those with competence in the field

Watching science activity : the coword maps visualize the structure of relationships between subjects of research and the way in which this network evolves with time. Thus, this method may be useful to identify subject research areas, and to investigate the distribution of publications, institutions, countries, in these areas of research. The goal is to indicate «who is doing what, where and when» (4W) with respect to the topics and centres of interest identified on the maps.

6. EPILOGUE

Today, the informetric techniques and the databases may be considered, in our opinion, as the contemporary instruments for representing and visualizing the state of scientific knowledge (natural and social sciences), the way Galileo turned the telescope on the heavens and set up the modern scientific revolution at the beginning of the Seventeenth Century.

Furthermore, we think that Price's instrumentality theory of innovation (see Price, 1984) can be applied to the informetric techniques field which offer new instrumentalities in order to produce a more empirical approach vis-à-vis traditional epistemology, taken to be the theory of scientific knowledge. As we know, Price coined the term instrumentality in order to indicate methods and techniques from which spring a scientific change or a new technology.

APPENDIX

Table 5. France

No	Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1	Judicial Organization	0.211	0.326	0.044	4	3	2	2	17	6
2	Criminality	0.070	0.279	0.052	7	15	5	1	21	14
3	Public Administration	0.099	0.241	0.058	10	13	1	5	36	21
4	Trade	0.208	0.236	0.005	4	3	2	0	20	6
5	Information	0.070	0.230	0.010	5	5	5	1	26	8
6	Teachers	0.052	0.183	0.007	6	6	2	0	27	18
7	Population	0.073	0.162	0.029	9	8	4	2	40	14
8	Attitude towards	0.032	0.159	0.019	8	7	10	5	69	18
9	Regions	0.020	0.158	0.015	6	5	10	5	65	6
10	Identity	0.021	0.154	0.042	10	13	8	5	41	9
11	Innovations	0.039	0.153	0.023	7	9	10	4	57	8
12	Art	0.127	0.151	0.038	4	3	1	1	15	9
13	Professions	0.100	0.135	0.051	10	13	4	9	59	25
14	Region	0.031	0.134	0.020	8	9	8	7	68	16
15	Industries	0.052	0.130	0.025	5	4	7	2	28	7
16	Social Class	0.052	0.129	0.017	4	3	2	1	18	6
17	Migrants	0.060	0.124	0.049	10	13	4	12	89	32
18	Youth	0.022	0.123	0.020	8	8	10	5	53	15
19	Enterprises	0.045	0.117	0.044	10	11	8	17	84	24
20	Religions	0.080	0.117	0.013	6	5	8	2	44	14
21	Transportation	0.059	0.115	0.034	10	16	1	8	64	32
22	Relations	0.036	0.109	0.030	10	10	6	28	176	31
23	Work Organization	0.076	0.107	0.042	8	10	8	11	73	12
24	Political Parties	0.030	0.105	0.018	10	11	7	8	110	42
25	Politics	0.030	0.105	0.024	10	11	8	10	101	33
26	Economic Sectors	0.018	0.094	0.025	10	10	10	12	88	11
27	Housing	0.044	0.093	0.019	10	9	6	2	61	28
28	Theory	0.027	0.051	0.012	4	3	10	2	56	28

Table 6. United Kingdom

No	Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1	Technological Innovation	0.100	0.284	0.020	5	6	6	2	16	4
2	Crime	0.067	0.224	0.052	9	13	6	4	29	15
3	State	0.032	0.205	0.010	4	3	5	0	25	14
4	Inequality	0.083	0.204	0.011	6	6	5	0	27	13
5	Work	0.032	0.184	0.029	8	10	9	10	65	16
6	Sanction	0.114	0.183	0.065	10	15	2	6	31	17
7	Ideologies	0.040	0.168	0.015	4	3	3	1	19	7
8	Careers	0.039	0.163	0.019	6	6	10	6	27	6
9	Gender Differences	0.018	0.138	0.022	8	7	10	7	59	17
10	Professions	0.041	0.135	0.012	6	6	4	0	27	11
11	Organisation	0.017	0.106	0.019	7	7	8	3	40	16
12	Family	0.021	0.102	0.028	9	8	8	4	44	11
13	Relations	0.042	0.088	0.032	10	10	7	35	128	44
14	Regions	0.044	0.084	0.024	4	3	5	3	19	6
15	Occupational Mobility	0.036	0.082	0.026	9	10	9	7	66	18
16	Enterprises	0.048	0.068	0.040	10	15	4	5	49	13
17	Females	0.040	0.057	0.028	6	6	9	17	63	5

Table 7. Germany

No	Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1	Generation	0.071	0.298	0.026	9	9	5	3	18	6
2	Woman's Work	0.044	0.270	0.038	10	12	8	4	18	6
3	Fertility	0.071	0.270	0.039	8	8	10	8	23	6
4	Application	0.108	0.247	0.072	9	14	5	24	52	11
5	Technology	0.095	0.243	0.082	7	15	2	11	21	3
6	Employment	0.045	0.229	0.023	9	8	10	4	26	5
7	Political Attitudes	0.222	0.222	0.005	4	3	1	0	6	4
8	Industry	0.100	0.208	0.043	7	7	10	4	18	3
9	Politics	0.038	0.201	0.038	10	12	8	6	26	3
10	Modernity	0.044	0.198	0.017	10	10	9	0	29	16
11	Social Policy	0.029	0.195	0.024	10	16	4	1	26	17
12	Social Movements	0.041	0.189	0.006	4	4	4	0	12	5
13	Structures	0.042	0.188	0.021	7	6	10	3	21	5
14	Research	0.050	0.169	0.023	5	4	10	5	17	3
15	Attitude towards	0.034	0.160	0.017	7	7	10	4	22	9
16	Social Change	0.045	0.158	0.048	10	10	8	12	42	10
17	Communication	0.057	0.157	0.015	4	5	3	0	10	7
18	Risk	0.036	0.145	0.034	9	11	9	4	28	8
19	Relations	0.033	0.143	0.040	8	9	10	48	71	8
20	Factors	0.074	0.142	0.048	10	10	9	8	22	4
21	Feminism	0.087	0.130	0.058	6	8	6	10	19	8
22	Work	0.054	0.125	0.031	7	6	9	9	23	2
23	Representation	0.063	0.122	0.007	6	5	5	0	16	6
24	Evolution	0.020	0.102	0.023	9	9	10	7	28	1

Table 8. United States

No	Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1	Mental Illness	0.200	0.263	0.000	5	4	0	0	14	8
2	Drugs	0.062	0.250	0.045	8	12	8	5	47	16
3	Blacks	0.045	0.225	0.015	6	7	7	4	39	9
4	Crime	0.111	0.215	0.054	7	8	8	11	84	16
5	Judicial Organization	0.061	0.203	0.060	10	15	5	11	85	29
6	Offenses	0.019	0.184	0.036	9	9	9	5	45	13
7	Sanction	0.134	0.168	0.090	9	14	3	17	53	8
8	Professions	0.024	0.150	0.010	6	6	4	1	30	10
9	Regions	0.028	0.146	0.013	8	8	9	3	71	33
10	Youth	0.045	0.136	0.032	5	4	10	5	50	10
11	Criminality	0.055	0.132	0.040	7	10	9	13	74	13
12	City	0.020	0.120	0.009	9	8	6	1	49	20
13	Police	0.047	0.116	0.027	5	5	10	4	55	12
14	Religious Attitudes	0.035	0.111	0.012	5	4	4	0	27	17
15	Family	0.049	0.098	0.035	7	6	9	9	67	16
16	Employment	0.036	0.077	0.015	7	7	10	2	65	24
17	Ethnic Groups	0.017	0.059	0.023	9	9	10	7	76	26
18	Relations	0.045	0.058	0.029	9	10	5	28	160	35
19	Females	0.030	0.049	0.024	10	13	5	8	107	39
20	Theory	0.027	0.043	0.019	10	9	4	1	104	68

REFERENCES

R.R. Braam, H.F. Moed, A.F.J. van Raan, “Comparison and Combination of Co-Citation and Co-Word Clustering”, in *Select Proceeding of the First International Workshop on Science and Technology Indicators*, Leiden, 14-16 November 1988, p. 307-337.

B.C. Brookes, “Information Space”, *The Canadian Journal of Information Science*, vol. 5, 1980, p. 199-211.

B.C. Brookes, “The Foundations of Information Science. Part IV: Information Science: The Changing Paradigm”, *Journal of Information Science*, vol. 3, 1981, p. 3-12

M. Callon, J-P. Courtial, W. A. Turner, S. Bauin, "From translations to problematic networks: An introduction to co-words analysis", *Social Science Information*, vol. 22, n° 2, 1983, p. 191-235.

M. Callon, J. Law and A. Rip (eds), *Mapping the Dynamics of Science and Technology*. London, Macmillan Press, 1986.

M. Callon, J-P. Courtial, F. Laville, “Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry”, *Scientometrics*, vol. 22, n° 1, 1991, p. 155-205.

J. Ducloy, P. Charpentier, C. François, L. Grivel, “Une boîte à outils pour le traitement de l’Information Scientifique et Technique”, *Génie logiciel*, n° 25, 1991, p. 80-90.

L. Grivel et J-Ch. Lamirel, “An analysis tool for scientometric studies integrated in an hypermedia environment”, in *Proceedings of ICO93* 4th International Conference on Cognitive and Computer Sciences for Organizations, Montreal, (Quebec) Canada, 4-7 mai 1993, p.146-154.

P. Healey, H. Rothman, P. Hoch, “An Experiment in Science Mapping for Research Planning”, *Research Policy*, vol. 15, 1986, p. 233-251.

P. Meincke and P. Atherton, “Knowledge Space: A Conveptual Basis for the Organization of Knowledge”, *Journal of the American Society for Information Science*, vol. 27, 1976, p. 18-24.

B. Michelet, L’analyse des associations. Paris: Thèse de doctorat, 1998.

A.J. Nederhof, R.A. Zwaan, R.E. de Bruin, P.J. Dekker, “Assessing the Usefulness of Bibliometric Indicator for the Humanities and the Social and Behavioural Sciences: A Comparative Study”, *Scientometrics*, vol. 15, n° 5-6, 1989, p. 423-433.

X. Polanco, “Analyse stratégique de l’information scientifique et technique. Construction de clusters de mots-clés”, *Sciences de la société*, n° 28, 1993, p. 111-126.

K.P. Popper, *Objective Knowledge.* Oxford: The Clarendon Press, 1979.

D. de S. Price, “Network of Scientific Papers”, *Science*, vol. 149, n° 3683, 1965, p.510-515.

D. de S. Price, "The Citation Cycle", p. 269 in *Little Science, Big Science ... and Beyond*. New York, Columbia University Press, 1986.

D. de S. Price, “The Science-Technology Relationship, the Craft of Experimental Science, and Policy for the improvement of High Technology Innovation”, *Research Policy*, vol. 13, 1984, p. 3-20.

H. Small and E. Garfield,"The Geography of Science: Disciplinary and National Mappings", in *Science Citation Index 1988*, Philadelphia: Institut for Scientific Information, p. 46-58.

W. Turner, G. Charton, F. Laville, B. Michelet, “Packinging Information for Peer review: New Co-word Analysis Techniques”, in A.F.J. van Raan (ed), *Handbook of Quantitative Studies of Science and Technology*. Amsterdam: Elsevier Science Publisher, 1988, p. 291-323.

P. H. Winston, *Artificial Intelligence.* London: Addison Wesley Publishing Co., 1977.

Chapitre 3[1]

Apports de l'analyse linguistique informatique dans l'analyse de l'information par la méthode des mots associés

Dès lors que l'on se propose de faire émerger le contenu cognitif d'un grand ensemble de documents et de le relier au contenu factuel (titres, noms d'auteurs, laboratoires, etc.), il peut être avantageux de s'appuyer sur des techniques linguistiques. Ici, les titres et résumés d'auteurs des notices bibliographiques, c'est-à-dire, des termes utilisés par les chercheurs eux-mêmes dans les documents scientifiques et techniques sont utilisés afin d'opérer une extraction terminologique et de s'affranchir de l'indexation manuelle pour éviter l'effet de l'indexeur. Comme son nom l'indique, cet effet désigne les conséquences du fait que l'indexation manuelle soit le produit de non-chercheurs, dont la formation scientifique serait en retard par rapport à la connaissance scientifique en action sur les fronts de la recherche.

L'objectif est de coupler les techniques linguistiques et infométriques afin de classifier et de représenter les connaissances véhiculées par les textes scientifiques et techniques sous leur forme écrite. Ce couplage doit *in fine* permettre de répondre à des questions stratégiques concernant beaucoup plus la connaissance que les documents eux-mêmes (informatique documentaire).

Les traitements linguistiques mis en œuvre reposent sur l'identification en corpus des termes d'une nomenclature terminologique (thésaurus, lexique d'indexation, glossaire, etc.) sous leurs formes de base ou sous des formes variantes. Ces traitements linguistiques améliorent sensiblement la collecte des termes comme le montre l'étude réalisée sur un corpus dans le domaine de la physique.

La technique des mots-associés a été appliquée pour mettre en évidence un réseau terminologique qui inclut à la fois des termes variants et non variants qui n'auraient pas été détectés sans ce traitement linguistique.

On peut repérer les clusters qui se singularisent par leur nombre élevé de termes variants. Il a été observé expérimentalement sur certains de ces clusters que cela correspondait à un changement d'activité (une accélération des recherches) dans le thème en question. Cependant, une certaine prudence s'impose : avant de qualifier ce phénomène d'indicateur d'activité scientifique, il faudrait le relier avec les indicateurs d'activité utilisés usuellement par les observatoires des sciences et technologie.

---

[1] Polanco X., Royauté J., Grivel L., Courgey A. 'Infométrie et linguistique informatique, une approche linguistico-infométrique au service de la veille scientifique et technologique', Les systèmes d'information élaborée, Ile Rousse, Corse, 1995.
Cet article est une version longue en français de l'article "How to do things with terms in informetrics : terminological variation and stabilization as science watch indicators"- présenté à la 5ème conférence internationale en infométrie, bibliométrie, scientométrie, River Forest USA, 1995.

1. Introduction.

Les travaux en informatique linguistique n'ont pas encore attiré beaucoup l'attention des chercheurs en infométrie. Ce texte fait part de nos récents travaux sur le couplage et l'application d'outils et de techniques en informatique linguistique et infométrie [12]. Par informatique linguistique, nous entendons tout traitement informatique du langage naturel qui permet de fournir des données linguistiques aux outils infométriques d'analyse de l'information scientifique et technique (IST).

Notre but est de construire un dispositif d'analyse de l'IST capable d'opérer à partir du texte intégral (ici, les titres et résumés d'auteurs des notices bibliographiques), c'est-à-dire, à partir des termes utilisés par les chercheurs eux-mêmes dans les documents scientifiques et techniques.

L'analyse repose sur des outils de classification automatique (SDOC et NDOC) utilisant les mots-clés (ou descripteurs) comme indicateurs de contenu [4]. Maintenant grâce aux traitements linguistiques, nous sommes capables de nous affranchir de l'indexation manuelle. Outre le fait qu'ils sont le point d'entrée du dispositif infométrique, les traitements linguistiques que nous effectuons ont aussi comme finalité de fournir des indicateurs infométriques pour la veille.

2. Objectifs et hypothèse.

Les objectifs que nous nous sommes fixés sont de trois types : technique, conceptuel et pragmatique. L'objectif technique est de coupler deux sortes d'outils : un outil scientométrique tel que le programme SDOC basé sur la technique des mots associés, et une plate-forme de traitement informatique du langage naturel. Rappelons ici que la méthode des mots associés a été proposée par M. Callon, J-P. Courtial et W. Turner pour la première fois au début des années quatre-vingt [1] [2] [3]. Quant à l'objectif conceptuel de notre approche, il est de classifier et de représenter les connaissances véhiculées par les textes scientifiques et techniques sous leur forme écrite, en nous appuyant sur les ressources de l'ingénierie linguistique et de la connaissance. La réalisation de cet objectif signifie un pas en avant dans notre projet d'une scientométrie cognitive [9]. Enfin, l'objectif pragmatique est de répondre à des questions stratégiques concernant beaucoup plus la connaissance que les documents eux-mêmes (informatique documentaire).

Les traitements linguistiques que nous mettons en œuvre reposent sur l'identification en corpus des termes d'une nomenclature terminologique (thésaurus, lexique d'indexation, glossaire, etc.), sous leurs formes de base ou sous des formes variantes. Nous considérons que, pour un corpus donné, le fait qu'un terme varie atteste que ce terme est "actif" puisqu'il est exprimé sous des formes traduisant des sous-aspects particuliers. A contrario, l'absence de variation peut être considérée comme un signe de stabilisation du concept dénoté par ce terme.

Nous faisons l'hypothèse que les phénomènes linguistiques de la variation et du figement des termes sont des indicateurs que l'on peut utiliser dans l'analyse des informations présentes dans le titre, dans le résumé, voire dans le texte même des documents scientifiques ou techniques.

3. Données, instruments et techniques.

Nous présentons dans cette section les instruments et les techniques de nature linguistique et infométrique que nous avons mis en place. Nous voulons dans un premier temps combiner ces deux types d'instruments, afin d'obtenir pour l'analyse infométrique des indicateurs linguistiques capables de représenter le contenu des documents (indicateurs de contenu), de manière plus complexe que les traditionnels mots-clés fournis par les notices bibliographiques elles-mêmes.

3.1. Données.

Nous détaillons ci-dessous les ressources documentaires nécessaires à l'expérience. Nous avons utilisé en premier lieu un thesaurus, le thesaurus du FIZ qui comporte 18 351 *master terms* (termes sous leurs formes préférentielles) et 2 804 *used-for* (synonymes). En second lieu, nous avons fait porter les traitements sur un ensemble de revues scientifiques : *Physical Review A, Physical Review B*, et *Applied Physics Letters* qui sont à l'origine de 519 références bibliographiques dans la base Pascal. Ces revues sont en anglais et comportent toutes des résumés. Les termes utilisés (au total 672) ont été extraits automatiquement des titres et des résumés d'auteurs de ces notices bibliographiques.

Il faut noter que *Physical Review A* est consacrée à la diffusion des travaux en physique atomique et moléculaire, tandis que *Physical Review B* et *Applied Physics Letters* diffusent les résultats de la recherche en physique de l'état condensé. Cette diversité doit se refléter au niveau des thèmes identifiés. Cela ne représente pas un problème dans la mesure où le but de l'expérience n'est pas d'analyser un domaine en particulier, mais de prouver l'importance et la faisabilité du projet que nous avons énoncé dans les sections précédentes (cf. §§ 1 et 2) et que nous détaillons par la suite.

3.2. Outil Infométrique.

Du point de vue infométrique, l'application de la méthode des mots associés (ici le programme SDOC) à l'ensemble des termes du titre et du résumé détectés par le traitement linguistique d'extraction terminologique, décrit ci-dessous (§ 3.3), nous a permis d'obtenir un réseau de termes variant peu ou figés, et de termes variant qui auraient été ignorés autrement. La variation offre la possibilité de capter les "signaux" faibles émis par ces termes et de les faire émerger. D'autre part, la classification a permis de situer ces termes dans des thèmes (au total, 20 clusters), qui se sont par ailleurs révélés être des structures complexes composées de pôles d'agrégation [12].

3.3. Outils linguistiques.

La chaîne linguistique-infométrique que nous avons mis en place s'appuie sur un analyseur (FASTR) [7] et sur un module d'assignation de catégories grammaticales (développé à l'INIST) pour l'étiquetage des mots de différents lexiques terminologiques [8] [13]. Ces outils permettent à partir d'une nomenclature terminologique quelconque, de repérer des termes sous leurs formes de base ou leurs formes variantes.

Nous identifions trois catégories de variations : 1) la variation flexionnelle, 2) la variation syntaxique et 3) la variation de type morpho-dérivationnelle. Chacune de ces variations pose un problème particulier pour la reconnaissance des termes. Nous mettons l'emphase sur la variation syntaxique qui est très productive. Les phénomènes de morphologie dérivationnelle sont cités à titre d'exemple et ne sont pas traités en tant que tels. Ils feront l'objet d'une étude ultérieure.

3.3.1. Variation flexionnelle.

Elle permet d'identifier pour chaque terme, les formes singulier / pluriel des noms (*deficiency : deficiencies*), et les formes infinitives, participe passées et gérondives des noms/verbes (*acoustic test : acoustic testing*). Dans les traitements que nous effectuons, chaque mot est décomposé en son lemme ou racine et sa terminaison. A chaque classe de mots correspond donc un lemme et ses différentes terminaisons.

3.3.2. Variation syntaxique.

La variation syntaxique est, avec la variation flexionnelle au centre des traitements que nous opérons. En effet, dans cette expérimentation, nous traitons trois sortes de variations syntaxiques :

(a) la *variation d'insertion* concerne tout mot à l'intérieur du groupe nominal, à l'exception de la plupart des mots grammaticaux. Par exemple, *X ray absorption spectroscopy* est associé au terme *X ray spectroscopy ;*
(b) la *variation de coordination* concerne toute forme coordonnéesde mots (adjectifs ou noms) à l'intérieur du groupe nominal. Par exemple, *differential and integrated cross sections* est associé au terme *Differential cross section* ;
(c) la *variation de permutation* implique tous les mots ou les groupes de mots pouvant permuter autour d'un élément pivot (prépositions ou séquences verbales). Par exemple, *range of power modulation frequency* est associé au terme *Frequency range*.

3.3.3. Variation morpho-dérivationnelle.

La variation morpho-dérivationnelle intègre dans la terminologie les phénomènes de nominalisation et d'adjectivisation. Ainsi la nominalisation de l'adjectif permet d'associer la séquence textuelle : *instable combustion* au terme *Combustion instability* ; dans les cas de nominalisation des verbes, "... *promotes degration of the cellular tumor...*" se trouve associé à *tumor promotion* et pour l'adjectivisation des noms : *optic disk* est équivalent au terme *optical disk*. En réalité, notre expérimentation ne traite pas les phénomènes de dérivation qui ne sont cités ici que pour l'exemple, elle ne traite comme nous avons dit que les variations flexionnelles (§ 3.3.1) et syntaxiques (§ 3.3.2).

4. Expérimentation.

Elle s'est déroulée en deux phases. La première, complètement automatique est le résultat brut du couplage du module d'extraction terminologique avec SDOC. Elle a permis d'obtenir, sans intervention humaine, une première classification. La deuxième phase a nécessité l'intervention d'un ingénieur documentaliste expert en physique.

La première étape de l'expertise a consisté à filtrer les termes du vocabulaire peu informatifs du point de vue du contenu. Il faut remarquer que la plupart des termes rejetés était des termes d'un seul mot (unitermes). Le bilan qui peut en être fait est que, outre la qualité finale de la classification, cette opération a été peu coûteuse en temps (un peu plus d'une demi-journée de travail d'expertise pour une personne). De plus, il est apparu qu'elle pourrait être automatisée de façon quasi-complète en la généralisant à l'ensemble du vocabulaire d'entrée.

Cette étape préalable d'épuration du vocabulaire a permis d'obtenir une classification en 20 thèmes principaux. Chacun de ces thèmes a pu être analysé et décrit par l'expert du domaine. Le tableau 1 donne ci-dessous le descriptif succinct de chacun de ces thèmes.

Ces thèmes ont été placés automatiquement sur une carte en fonction des critères de cohésion et de centralité propres à l'outil d'analyse infométrique SDOC. La cohésion caractérise la valeur des associations unissant les mots qui composent un thème donné. La centralité rend compte pour un thème de la valeur de ses associations avec d'autres thèmes. Ces deux mesures permettent de ranger les différents thèmes sur un plan bidimensionnel (voir figure 1) .

D'une façon générale, la carte est un indicateur de l'importance relative des thèmes par rapport à la cohésion (y) de l'information qu'ils représentent individuellement, et à la centralité (x), c'est-à-dire le rôle qu'ils jouent dans l'ensemble du domaine au moment de l'analyse. C'est aussi un moyen de représenter automatiquement les contenus de connaissance véhiculés par les documents à partir de la terminologie utilisée dans un cluster (agrégat ou amas). En résumé, ce type de carte est un outil d'aide à l'analyse de l'information.

OPTICAL PROPERTIES	Propriétés optiques - Lasers
SIZE	Effets dimensionnels - Super-réseaux
COUPLINGS	Diffusion des impuretés (en particulier l'hydrogène). Interaction particules-particules et particules-rayonnement
IONIZATION	Ionisation, transitions électroniques dans les atomes, les molécules et la matière condensée
ELECTRON DENSITY	Etudes de la densité électronique et des ondes de densité de charge
ELECTRIC FIELDS	Champs électriques (influence, comportement) dans les atomes, molécules et dans la matière condensée
SCATTERING	Phénomènes de transport
THIN FILM	Structures, propriétés des couches minces et des monocristaux
HETEROSTRUCTURES	Puits quantiques et autres hétérostructures
VALENCE	Phénomènes relatifs à la structure électronique dans les semiconducteurs
PHOTOLUMINESCENCE	Etudes de la photoluminescence sur des couches semiconductrices ou des puits quantiques
INTERACTIONS	Interaction entre particules ou quasiparticules
LAYERS	Croissance et dépôt de couches minces
ELECTRONIC STATES	Structure et phénomènes électroniques dans la matière
SURFACES	Etats électroniques et phénomènes électroniques de surface
MAGNETIC FIELDS	Influence d'un champ magnétique sur la matière condensée
PHOTONS	Interaction des atomes et des molécules avec un rayonnement électromagnétique
IRRADIATION	Etude des phénomènes dus à une irradiation par particules ou rayonnement électromagnétique
QUANTUM WELLS	Puits quantiques, barrières de potentiel, confinement optique
GROWTH	Croissance et dépôt de couches minces

Tableau 1 — Descriptif des thèmes

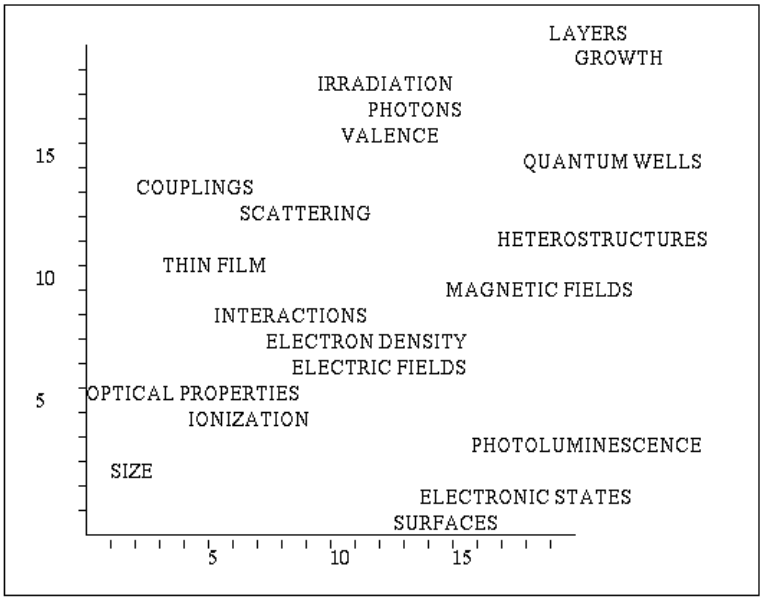


Figure 1 — Carte global du domaine analysé

Il n'est pas question d'entrer ici dans des considérations plus détaillées sur la technique de construction de la carte et son mode d'emploi dans l'analyse de l'IST. Nous l'avons déjà fait à plusieurs reprises [10] [11], ainsi que nos collègues qui sont à l'origine de la méthode des mots associés [1] [2] [3]. Nous nous contenterons ici de souligner que, dans notre dispositif, la carte joue le rôle d'une surface de représentation de ces structures complexes que sont les clusters. Comme il a été expliqué (§ 3), l'emploi d'un outil de classification automatique nous permet de replacer les phénomènes linguistiques de la variation et du figement (cf. § 3.2) à l'intérieur des clusters, une fois que ces phénomènes ont été automatiquement repérés au niveau des textes eux-mêmes par les traitements informatiques linguistiques (cf. § 3.3). On verra dans la section 5 comment ces phénomènes de langue sont représentés d'une manière quantitative (suite à un calcul) et qualitative dans la structure des clusters.

5. Discussion.

Le but de cette section est de montrer l'apport que signifie la mise en œuvre de notre hypothèse (cf. § 2); à savoir que la variation et le figement peuvent être des indicateurs linguistiques de connaissance, susceptibles d'être l'objet d'une mesure dans le cadre d'une ingénierie de l'information, et donc exploitables du point de vue infométrique. Il s'agit ici de mesurer le rôle qu'ils jouent dans les clusters et les types de clusters qu'ils privilégient. Il faut noter que les clusters sont des indicateurs des thèmes ou des centres d'intérêt autour desquels s'agrège l'information à un moment donné.

Quand on observe les termes qui ont servis à créer les clusters, on remarque que certains varient beaucoup alors que d'autres sont remarquables par leur stabilité. Que traduit ce phénomène de langue et comment lui donner une interprétation en terme d'indicateurs de connaissance ? Nous cherchons dans un premier temps à donner une explication linguistique à ces observations (§ 5.1). Ensuite, nous voulons en donner une mesure (§ 5.2), afin de pouvoir les utiliser comme indicateurs et les appliquer dans notre approche infométrique (§ 5.3). Enfin, nous tentons de réanalyser les clusters du point de vue de la variation et du figement (§§ 5.4 ; 5.5).

5.1 Variation et figement.

Si l'on considère les termes d'un sous-domaine quelconque comme un sous-ensemble particulier des noms composés (*carte bleue*, *ceinture noire*, *homme grenouille* pour la langue courante ; *champ magnétique* / *magnetic field*, *niveau de Fermi* / *Fermi level*, *potentiel électrique* / *electric potential* pour les termes de physique), nous sommes alors confrontés à la problématique du figement. Il s'agit d'une notion importante d'un point de vue sémantique, car elle confère au terme une valeur référentielle relativement stabilisée par rapport au concept, valeur qui est partagée par une communauté professionnelle (ici les physiciens). Il faut préciser qu'en cas de figement le sens du terme n'est pas directement déductible de la composition du sens des mots qui le forme. Remarquons que le critère du figement ne fait pas l'unanimité dans la communauté linguistique (voir le panorama sur la question dans [6] et le numéro spécial de TAL [14] consacré à ce sujet). Nous en donnerons une définition simple, qui, sans être complète du

point de vue linguistique, a surtout le mérite d'être opératoire par rapport à ce que nous sommes capables d'observer.

On considère comme figé tout terme pour lequel les éléments qui le composent sont indissociables, et pour lesquels l'ordre et la contiguïté de ses éléments sont stables ou faiblement affectés.

Cette définition exclut les groupes nominaux ordinaires pour lesquels on n'observe pas ce type de contrainte, à part les contraintes de bonne formation du syntagme. Il a été montré que le figement n'est pas un critère absolu, mais qu'il existe des degrés de figement reposant sur des propriétés transformationnelles propres au groupe nominal [5]. Les variations d'insertion, de coordination et de permutation sont les "opérations" de notre définition et elles reposent sur ces propriétés transformationnelles. Sans entrer plus dans le détail, nous considérerons, conformément à notre définition, comme plus figé un terme qui n'admet pas l'insertion, la coordination ou la permutation, qu'un terme qui les accepte (exemple: le terme *Electron collisions* soumis à la variation d'insertion *electron molecule collision* ; de permutation : *collision strengths for electron* ; ou de coordination : *electron and hole collisions*).

Si l'on relie variation et figement, la variation est ce que l'on peut observer pour un terme t dans un corpus C, le figement est ce que l'on peut éventuellement déduire de cette observation, car ce n'est pas parce qu'aucune variation n'est constatée pour le terme t, que celui-ci est figé. Les tests linguistiques de figement imaginés par G. Gross [5] pourraient donner une indication fiable, mais ils ont l'inconvénient de ne pas être automatisables, et de nécessiter une double expertise (celle du linguiste et celle du spécialiste du domaine analysé). Pour ces raisons, il nous a semblé utile d'en donner une approximation à partir des données du traitement automatisé. Nous considérerons comme un indice du figement d'un terme le nombre réduit de formes variantes de ce terme ou leur absence, par rapport aux formes de base observées. Autrement dit, nous interprétons le faible emploi de formes variantes d'un terme donné, comme le signe manifeste du figement de ce terme dans l'usage.

Ces réflexions nous ont permis de formuler l'hypothèse que la variation et le figement peuvent être des indicateurs de connaissance que l'on peut mesurer par l'affectation d'un poids. Nous avons donc créé deux indicateurs : VAR_i pour la variation, et FIG_i qui reflète les potentialités d'un terme à être figé.

5.2 Indicateurs de variation et de figement.

L'observation des données montre que variation et figement ne sont pas des phénomènes symétriques. La variation d'un terme est toujours associée à un nombre important d'occurrences de ce terme sous sa forme de base. On appelle forme de base celle qui est enregistrée dans la nomenclature du domaine. Le figement correspond à une minimisation des formes variantes du terme (tendant vers 0) par rapport à la forme de base, et il ne peut pas être admis de parler de figement si les formes variantes sont plus nombreuses que les formes de base. Ce critère de minimisation n'existe pas pour la quantification de la variation et il n'est pas absurde de lui donner une valeur si les formes de base sont plus nombreuses que les variantes.

Soit f_{ij} un entier qui prend la valeur 1 quand il existe une ou plusieurs variations du terme i dans le document j ; T le nombre de documents du corpus. Alors, n, le nombre de documents comportant des variations du terme i est égal à _ f_{ij}. Soit N, le nombre de documents indexés par le terme i ; alors (N - n) est le nombre de documents indexés par la forme normale du terme i. On désigne par VAR_i, l'indice de variation du terme i et par FIG_i, l'indice de figement du terme i.

Nous proposons un indice de variation qui privilégie les termes qui varient beaucoup dans le plus grand nombre de documents :

$$\mathbf{VAR_i = (n^2 / N) / T = n^2 / N * T} \qquad \mathbf{(1)}$$

VAR_i tend vers 1 pour tout terme apparaissant au moins une fois dans chaque document sous une forme uniquement variante (pour n = N = T).

L'indice de figement privilégie les termes variant peu ou pas dans le plus grand nombre de documents ; $\Delta = (N - 2n)$ est la différence entre le nombre de documents où le terme apparaît sous sa forme de base et le nombre de documents où il est sous une forme variante ; et (N -n) est le nombre de documents où un terme donné apparaît sous sa forme de base. Cela donne la formule suivante :

$$\mathbf{FIG_i = D * ((N-n) / N) / T = D * (N-n) / N * T} \qquad \mathbf{(2)}$$

FIG_i est significatif seulement pour $\Delta > 0$.

FIG_i tend vers 1, pour tout terme apparaissant au moins une fois dans chaque document sous une forme non variante (pour n = 0 et N = T).

En donnant à ces phénomènes linguistiques une expression quantitative, nous produisons une nouveau type d'indicateurs. En effet, nous sommes partis de l'hypothèse que la variation, mais aussi l'absence de variation, pouvaient être utilisées à des fins de veille scientifique. Nous disposons maintenant d'indices permettant de les mesurer afin de les interpréter.

5.3. Application.

Les tableaux 2 et 3 présentent ci-dessous une liste de termes parmi les plus significatifs classés à l'aide de ces deux indicateurs, VAR_i et FIG_i (multipliés par 1000 pour une meilleure lisibilité). Ainsi nous pouvons observer quel rôle jouent ces termes dans les clusters et quels types de clusters ils privilégient.

Les termes les plus figés (termes complexes de plus de deux mots) ont la particularité de se répartir dans des clusters différents (10 termes sur les 13 du tableau 2), plutôt que de se regrouper dans un ou deux clusters significatifs. Quand on regarde l'ensemble des clusters, il y a toujours au moins un terme fortement figé. Les termes les plus figés ne permettent pas de différencier les clusters, mais en tant que "signal fort" (entre 12 à 42 occurrences pour le tableau 2) ils participent activement au processus de classification.

Quand on examine les termes les plus variants (tableau 3), on remarque qu'un nombre important parmi eux (21 termes sur les 35 les plus sujets à variation) n'appartiennent à aucun cluster. La classification n'a pas permis de capter tous les termes significatifs de ce phénomène de langue. Cela est dû en partie au seuil de cooccurrence fixé dans ce cas à 3 cooccurrences, en vue d'obtenir un nombre réduit de classe, mais qui a le désavantage de rejeter certains de ces termes.. Dans nos prochaines expérimentations, nous donnerons un poids plus grand aux termes variants, afin qu'aucun de ces termes ne puissent être rejetés du processus de classification.

Termes	N	n	FIG_i
QUANTUM WELLS	42	1	75.24
MAGNETIC FIELDS	37	1	65.61
GROUND STATES	26	0	50.10
CROSS SECTIONS	23	0	44.32
ELECTRICAL FIELDS	23	1	38.70
FERMI LEVEL	19	0	36.61
ELECTRIC POTENTIAL	19	0	36.61
THIN FILMS	16	0	30.83
MOLECULAR BEAMS	15	0	28.90
EFFECTIVE MASS	15	0	28.90
ENERGY LEVEL DENSITY	16	1	25.29
BANS STRUCTURE	16	1	25.29
MOLECULAR BEAM EPITAXY	12	0	23.12

Tableau 2 — Les termes les plus figés (échantillon).

Les 35 termes les plus variants se regroupent dans 6 clusters : IRRADIATION, ELECTRIC FIELD, ELECTRONIC DENSITY, PHOTONS, SURFACES et VISIBLES RADIATION. Ces termes les plus variants dans les clusters sont surtout liés aux aspects des champs électriques dans la matière condensée, rayonnement électromagnétique, phénomènes électroniques de surface et lasers.

5.4. Les clusters et les phénomènes de variation et de figement

La variation n'est pas un phénomène numériquement important et, pour l'observer, il faut un nombre significatif de documents. Quand on regarde les textes, on ne trouve pas de résumé qui se singularise du point de vue de ce phénomène. S'il n'existe pas de texte singulier du point de vue de la variation, qui est un signal trop faible pour être détecté dans des courts résumés d'auteurs, il fallait trouver un moyen de l'observer et de l'opposer au figement. Les clusters semblaient être le lieu logique d'une telle observation.

Termes	Variantes	Variations
SURFACE ENERGY		
	energies of si surfaces	Perm
N = 8	energy dissipation in sliding crystal surfaces	Perm
n = 6	energy for a number of surfaces	Perm
VARi = 8,67	energy necessary to achieve a given surface	Perm
	surface free energies	Ins
	surface state energies	Ins
	surface state energy	Ins
X-RAY SPECTRA		
	spectra produced by x ray	Perm
N = 6	x ray absorption spectrum	Ins
n = 5	x ray emission spectra	Ins
VARi = 8,03	x ray photoemission spectrum	Ins
	x ray scattering spectra	Ins
FIELD IONIZATION		
	field induced ionization	Ins
N = 6	field multiphoton ionization	Ins
n = 5	ionization by strong fields	Perm
VARi = 8,03	ionization in strong laser fields	Perm
	ionization in very intense radiation fields	Perm
	ionization probability decreases with increasing field	Perm
ELECTRON COLLISIONS		
	collision strengths for electron	Perm
N = 6	electron and hole collisions	Coor
n = 5	electron atom ionizing collisions	Ins
VARi = 8,03	electron h2 collisions	Ins
	electron molecule collision	Ins
EXTERNAL FIELDS		
	external bias field	Ins
N = 9	external electric field	Ins
n = 6	external magnetic field	Ins
VARi = 7,71	external magnetic fields	Ins
SURFACE PROPERTIES		
N = 4	properties of a lateral surface	Perm
n = 3	properties of lateral surface	Perm
VARi = 4,34	properties of the al surfaces	Perm
EPITAXIAL LAYERS		
N = 5	epitaxial insulating layer	Ins
n = 3	epitaxial si1 xGex layers	Ins
VARi = 3,47	epitaxial siC conversion layer	Ins
PULSED LASERS		
	laser ablation to produce a pulsed	Perm
N = 9	pulsed and cw laser	Coor
n = 4	pulsed ruby laser	Ins
VARi = 3,43	pulsed xeCl laser	Ins

Tableau 3 — Les termes les plus variants (échantillon).

Si les phénomènes de variation et de figement peuvent s'interpréter en termes d'indicateurs de connaissance, ils doivent nécessairement trouver une expression dans les clusters, même si la classification a tendance à rejeter un nombre important de termes variants. Afin de mettre en évidence l'effet de la variation et du figement au sein des clusters, nous avons été amenés à imaginer une façon de les classer, en prenant en compte le fait qu'un cluster est constitué à la fois de termes simples (un seul mot) et de termes complexes (plusieurs mots). Nous avons donc retenu pour ce classement le coefficient de variation (VAR_i) qui assigne la valeur nulle aux termes ne variant pas. Il était donc naturel de considérer les termes simples (formés d'un seul mot) tout autant stabilisé que les termes complexes (formés de plusieurs mots) ne variant pas. Ainsi nous avons attribué la valeur nulle aux termes d'un seul mot.

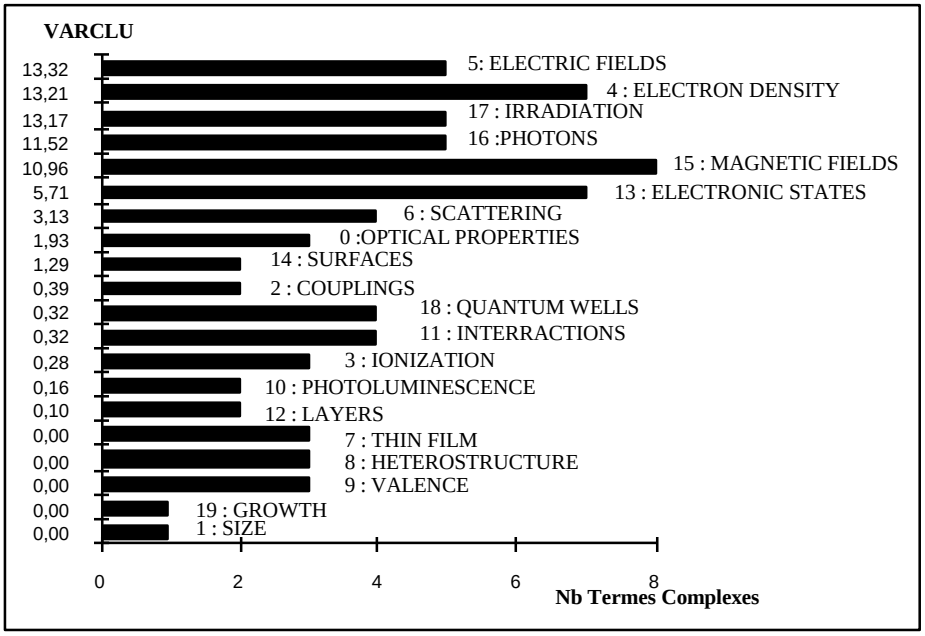


Tableau 4 — La variation dans les clusters.

L'indice de variation du cluster (VAR_{CLU}) est obtenu par un calcul simple qui consiste à sommer, pour chaque cluster, les différents coefficients de variation de chacun des termes complexes qui le composent :

$$\mathbf{VAR_{CLU}} = \mathrm{S}\mathbf{VAR_i}$$

Dans le tableau 4, les clusters sont ordonnés par valeurs croissantes de leur indice VAR_{CLU}. Ce classement des clusters selon le critère de la variation des termes qui le composent (VAR_{CLU}) permet d'opposer deux ensembles de clusters :

(a) Ceux avec les termes les plus figés : SIZE, GROWTH, VALENCE, HETEROSTRUCTURE, THIN FILMS, LAYERS, PHOTOLUMINESCENCE. Ces clusters sont liés plus particulièrement à des thématiques concernant les couches minces, leurs croissances et dépôts, les phénomènes électroniques dans les semi-conducteurs et la photoluminescence sur les couches semiconductrices, les hétérostructures.

(b) Ceux avec les termes les plus variants : ELECTRIC FIELDS, ELECTRON DENSITY, IRRADIATION, PHOTONS, MAGNETIC FIELDS et ELECTRONIC STATES. Ces clusters concernent les champs électriques et l'influence des champs magnétiques sur la matière condensée, la densité électronique et les phénomènes électroniques dans la matière, le rayonnement électromagnétique et les interactions atomes / molécules.

A partir de cette observation, deux clusters (GROWTH et ELECTRONIC STATE) représentatifs de ces deux ensembles ont été choisis, afin de les analyser en détail du point de vue de ce qu'un expert du domaine pouvait observer quant aux rôles de ces phénomènes de langue au niveau des clusters (voir les graphes des figures 3 et 4 relatifs à

ces deux clusters) . Les remarques de la section suivante sont principalement le résultat de ce travail d'expertise.

5.5 Analyse de deux thèmes représentatifs de la variation et du figement

Il est nécessaire, auparavant, de rappeler les propriétés de la méthode infométrique que nous utilisons. Identifier les clusters et décrire les associations qui les constituent (intra-clusters) et qui les unissent (inter-clusters) représentent la première étape dans l'analyse de l'information. Ensuite, il s'agit de caractériser la structure d'ensemble du réseau et la contribution de chacun des clusters (thèmes) à sa structuration. Ainsi, les notions de centralité et de cohésion (ou densité) sont destinées à mettre en évidence la contribution des différents clusters (agrégats ou amas) à la structuration du réseau global (figure 1).

La centralité (sur l'abscisse) mesure pour un cluster la force de ses associations avec d'autres clusters (relations inter-clusters). Plus ces associations sont nombreuses et fortes, plus le cluster désigne un ensemble de problèmes de recherche d'importance dans l'ensemble de l'information scientifique et technique que l'on analyse.

La cohésion ou densité (sur l'ordonnée) mesure la force des associations qui unissent les mots qui composent un cluster. Plus ces associations sont fortes et plus les problèmes de recherche correspondant au cluster constituent un ensemble cohérent et intégré. Quand ces associations intra-cluster sont faibles, le cluster présente une structure interne molle, éclatée, ce que l'on peut interpréter comme l'indice d'un thème constitué par des unités d'information relativement désagrégées. Même si, comme l'a constaté l'expert du domaine, le cluster ELECTRONIC STATES est homogène, il a pourtant une valeur de cohésion très faible, comme nous pouvons l'observer sur la carte (voir les figures 1 et 2 et le graphe de la figure 4).

Outre les informations fournies par la méthode des mots associés relatives aux propriétés des clusters, nous disposons maintenant des informations linguistiques telles que la variation et le figement concernant les termes qui composent les clusters. L'expertise a consisté à faire une lecture des clusters du point de vue du contenu scientifique qu'ils représentent, en exploitant toutes ces informations.

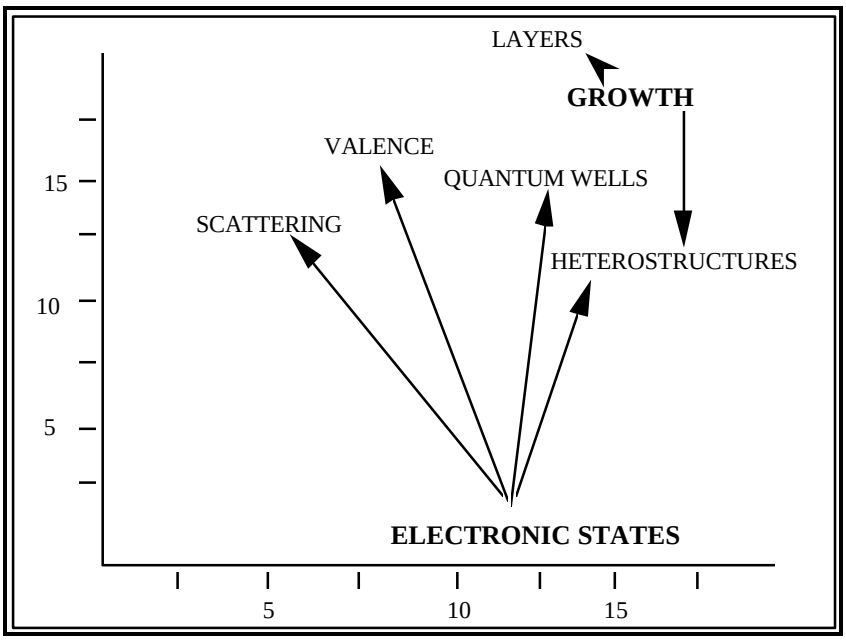


Figure 2 — Carte locale représentant les deux thèmes choisis : GROWTH et ELECTRONIC STATES et leurs associations (externes) avec d'autres thèmes (réseaux locaux). On voit que ces deux thèmes sont reliés à travers le thème HETEROSTRUCTURES. Leurs positions sur la carte montrent que si ils sont relativement proches sur l'axe de la centralité (x), mais assez distants sur l'axe de la cohésion (y).

GROWTH — Ce thème fait partie des thèmes composés de termes variant peu. Il regroupe 45 articles qui traitent de la croissance et du dépôt de couches minces. Les mots-clés du thème évoquent soit les couches minces elles-mêmes (FILMS, MONOLAYERS, MULTILAYERS, LAYERS), soit le phénomène de croissance de la couche (GROWTH, ISLANDS, NUCLEATION), soit la méthode de dépôt utilisée (DEPOSITION, VAPORS, CVD, PLASMA, MOLECULAR BEAMS). Enfin le mot-clé GRAPHITE se rapporte à un support fréquemment utilisé pour le dépôt de couches minces. Ce thème est relié (à travers les associations externes fondées sur la cooccurrence des mots qui se sont agrégés dans des clusters différents) aux thèmes LAYERS (couches) et HETEROSTRUCTURES (constituées par une superposition de couches).

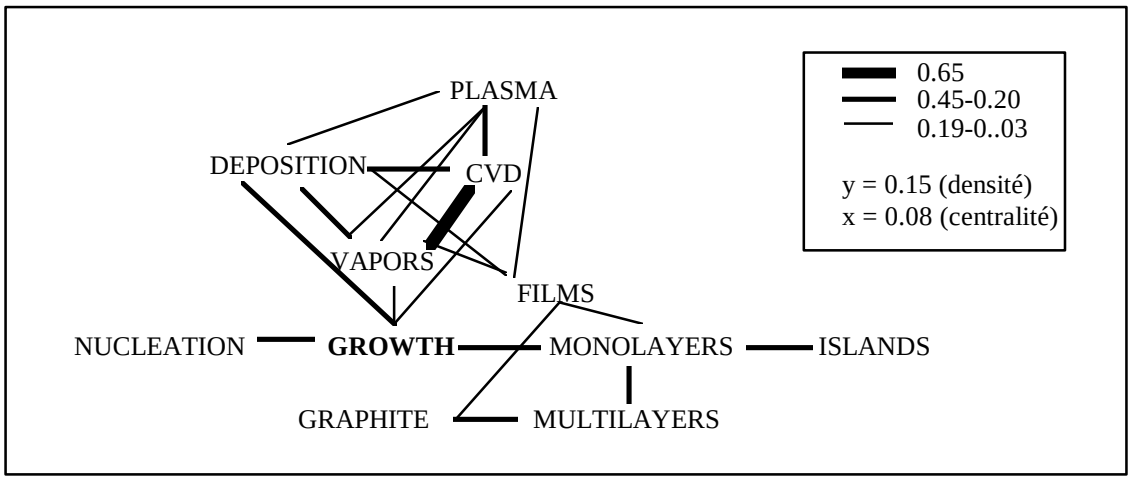


Figure 3 — Graphe représentant les associations (internes) entre les termes composants du clusters GROWTH. La valeur plus élevée de son indice de densité ou de cohésion interne (y) explique sa position en haut de la carte (figures 1 et 2). On voit ici un cluster qui présente une structure forte, à cause justement de la valeur moyenne de ses associations internes.

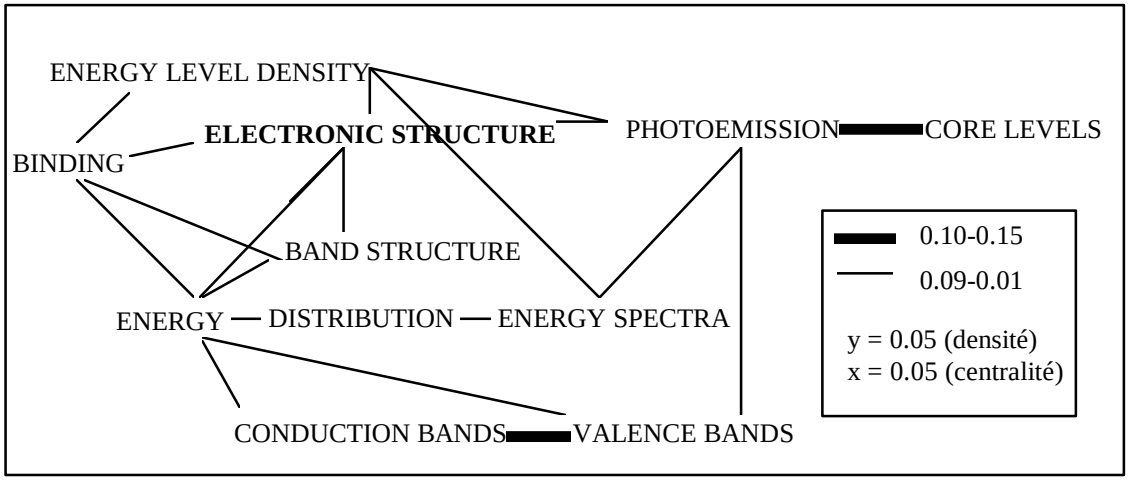


Figure 4 — Graphe représentant les associations (internes) entre les termes composants du cluster ELECTRONIC STATE. Il faut noter que ce cluster a été renommé par l'expert du domaine d'un point de vue conceptuel. La faible valeur de son indice de densité ou de cohésion interne (y) explique sa position en bas de la carte (figures 1 et 2). Il s'agit d'un cluster qui présente une structure molle, à cause justement de la valeur faible de ses associations internes.

ELECTRONIC STATE — Ce thème rassemble des articles traitant de la structure électronique (niveaux d'énergie, structure de bande) et des phénomènes électroniques (propriétés optiques, photoémission) dans la matière condensée (incluant les puits quantiques et autres hétérostructures). C'est un thème très homogène dont tous les mots-clés qu'ils soient obtenus par associations internes ou externes, évoquent le comportement des électrons dans la matière : ENERGY est en effet une composante de l'expression ENERGY-LEVEL, relative aux niveaux d'énergie électronique.

On observe pour ce thème de nombreux phénomènes de variation que l'on peut classer en différents types :

(a) Certaines variations introduisent un autre objet ou un autre phénomène que celui contenu dans le mot-clé original. C'est souvent le cas avec la variation de coordination : le terme CONDUCTION BANDS est observé dans l'expression : *conduction and valence band*. C'est ainsi que dans cette expression est introduite une autre bande qui est la bande de valence. On rencontre le même phénomène pour le terme VALENCE BANDS.

(b) D'autres variations (les plus nombreuses) apportent une précision sur l'objet ou le phénomène décrit. Par exemple le terme ELECTRONIC STRUCTURE est obtenu par renvoi de synonymie du thesaurus sur le terme ATOMIC SHELLS. Certaines

variations de permutation combinées avec l'insertion (*Auger spectrum*, la séquence insérée) vont permettre de repérer ce synonyme dans les expressions : *shell Auger spectrum in atomic*, *shell Auger spectrum of atomic*, précisant le type d'étude (spectre Auger) de la structure électronique réalisée. Précisons toutefois que cette variation aurait dû être rejetée avec des métarègles plus filtrantes (les métarègles sont des règles qui permettent de contrôler le processus de repérage des termes variants), car *atomic* est un adjectif qui ne peut spécifier que le nom qui suit et qui ne peut donc pas permuter (du point de vue stricte de la syntaxe). Cependant, d'un point de vue sémantique, le terme identifié est valide.

Les expressions *electronic band structure*, *electronic subband structure* (obtenues par la variation d'insertion) précisent que l'on a affaire à une structure de bandes, et l'expression *density of inoccupied states* apporte une précision (*inoccupied*) sur les états électroniques décrits.

Le terme ENERGY SPECTRA, repéré également à l'aide de l'insertion, apparaît dans les expressions : *energy Auger spectrum*, *energy loss spectra*, *energy résolution photoemission spectra*, qui, toutes, précisent le type de spectre décrit.

(c) Une autre variation observée n'apporte pas vraiment de précision sur l'objet ou le phénomène décrit. C'est ainsi que l'on trouve le terme CORE LEVELS sous la forme *core electron levels*, ce qui dans le contexte de la photoluminescence n'apporte rien de plus que *core levels*.

(d) Enfin, certaines variations sont dues au fait que l'auteur fait référence à ce dont il vient de parler : *spectra in this energy*, *structure on the latter band*.

L'analyse de ce dernier thème montre que la variation permet de rendre l'analyse plus précise et plus fine. Du point de vue de l'analyse des thèmes, nous vérifions ci-dessus que pour la variation d'insertion, chaque élément inséré est porteur d'une information de contexte utile à exploiter. Cette information de contexte, nous la retrouvons dans la permutation quand celle-ci se compose avec la variation d'insertion : *properties of lateral surfaces* lié au terme SURFACE PROPERTIES, où l'adjectif *lateral* spécifie *surface*. La coordination apporte une information de proximité sémantique entre deux termes : la séquence *conduction and valence band* montre que les termes CONDUCTION BANDS et VALENCE BANDS peuvent se coordonner parce que sémantiquement proche. Cette proximité sémantique est vérifiée également par le lien de cooccurrence des deux termes dans le cluster.

6. Conclusion.

1. Maintenant, il importe de s'interroger sur le nouvel objet que nous avons créé par le couplage des outils infométriques et linguistiques ; à savoir, un réseau de termes dont certains varient fortement et d'autres remarquables par leur stabilité non variationnelle. C'est ce réseau qui à l'avenir devra être interrogé, à partir de ces indicateurs, en nous permettant de signaler et de mesurer des phénomènes de stabilité ou d'instabilité au niveau des termes employés dans les textes scientifiques ou techniques.

2. Le fait est que nous disposons désormais d'un instrument linguistico-infométrique permettant la visualisation des informations présentes dans les titres, les résumés, voire dans le texte, et qui comporte si l'on peut ainsi s'exprimer, trois niveaux successifs de résolution : le niveau *macro*, c'est-à-dire la carte de clusters ; le niveau *meso* qui est représenté par les clusters eux-mêmes ; et enfin, le niveau *micro*, autrement dit le réseau de termes avec leurs variations et leurs absences de variation syntaxique.

3. Quant à l'hypothèse que les phénomènes linguistiques de la variation et du figement peuvent être des indicateurs de connaissances (c'est-à-dire de la connaissance écrite véhiculée par les textes scientifiques et techniques), l'expérience réalisée soulève trois remarques.

   (a) En raison des paramètres, la classification a rejeté un nombre non négligeable des termes variants, à cause justement de leur cooccurrence faible (dans ce cas concret, inférieure à trois); il nous faut donc tenir compte de ce phénomène et trouver le moyen de le corriger. Dans nos prochaines expérimentations, nous envisageons de donner un poids plus grand aux termes variants, afin qu'aucun de ces termes ne puissent être rejeté du processus de classification.

   (b) La seconde remarque est qu'on a pu contraster deux ensembles de clusters suivant le critère de la variation des termes qui les composent (§ 5.3) ; du point de vue du contenu scientifique, est apparu pour l'expert du domaine qu'une telle distribution (cf. tableau 3) correspondait à la distinction qu'il pouvait reconnaître entre thèmes d'ordre plus théorique (VAR_{CLU} important) et thèmes d'ordre beaucoup plus applicatif (VAR_{CLU} faible). Pourtant, nous devons nous garder de tirer des conclusions un peu hâtives visant à associer termes variants et langage théorique (ou problèmes théoriques de recherche), et termes figés et langage applicatif (ou problèmes d'application).

   (c) Et pour conclure, notre dernière remarque est qu'il apparaît nécessaire, pour véritablement tester le rôle d'indicateurs de ces phénomènes de langue, de travailler sur des ensembles plus importants de termes et, par là même, à partir d'ensembles plus importants de textes pleins.

7. Références.

[1] Callon, M., J. Law, A. Rip (1986), *Mapping the Dynamics of Science and Technology*. London, MacMillan.

[2] Callon, M., J-P. Courtial et H. Penan (1993), *La scientométrie*. Paris: Presses Universitaires de France. Que sais-je?, N° 2727.

[3] Courtial, J-P.(1990), *Introduction à la scientométrie*. Paris, Anthropos-Economica.

[4] Grivel, L. et C. François (1995), "Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique", *SOLARIS*, n° 2, à paraître.

[5] Gross, G. (1988), "Structure des noms composés", *Informatique & Langue Naturelle*, ILN'88, Nantes, France. Octobre

[6] Habert, B. et C. Jacquemin, “Noms composés, termes, dénominations complexes : problématiques linguistiques et traitement automatiques”, *Traitement Automatique des Langues*, 34 (2),1993, p. 5-42.

[7] Jacquemin, C. (1994), "FASTR: A Unification-based Front-end to Automatic Indexing", *RIAO 94 Conference Proceedings «Intelligent Multimedia Information Retrieval Systems and Management»*, Rockfeller University, New York, October 11-13, p. 34-47.

[8] Jacquemin, C., et J. Royauté (1994), "Retrieving Terms and their Variants in a Lexicalised Unification-Based Framework", *Proceedings 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3 - 6 July, Dublin.

[9] Polanco, X., L. Grivel, C. François et D. Besagni (1993), "L'infométrie, un programme de recherche", *Journées d'études "Les systèmes d'information élaborée"*. Ile Rousse, Corse, France, 9-11 Juin, texte n° 3.

[10] Polanco, X. (1993), "Analyse de l'information scientifique et technique. Construction de clusters de mots-clés", *Sciences de la société*, n° 29, p. 111-126.

[11] Polanco, X. et L. Grivel (1993), "Mapping Knowledge: The Use of Co-Word Analysis Techniques for Mapping a Sociology Data File of Four Publishing Countries (France, Germany, United Kingdom and United States of America)”, *Fourth International Conference on Bibliometrics, Informetrics and Scientometrics.* 13-18 September, Berlin, Germany.

[12] Polanco, X., L. Grivel, J. Royauté, “How to Do Things with Terms in Informetrics: Terminological Variation and Stabilization as Science Watch Indicators”, *Fifth International Conference on Scientometrics & Informetrics*, River-Forest (Chicago), Illinois, USA, June 7-10, 1995, à paraître.

[13] Royauté, J. et C. Jacquemin (1993), "Indexation automatique et recherche de noms composés sous leurs différentes variations". *Informatique & Langue Naturelle*, ILN'93, Nantes, France. Décembre

[14] *Traitement Automatique des Langues* 34 (2), 1993, Revue de l’Association pour le Traitement Automatique des Langues (ISSN 0039-8217).

Chapitre 4[1]

Génération automatique d'hypertextes avec cartes thématiques : avant le World Wide Web

Ce chapitre approfondit la démarche d'analyse ébauchée dans les chapitres précédents, en montrant plus particulièrement comment l'utilisation traditionnelle du diagramme stratégique dans la méthode des mots associés peut être complétée par une analyse des relations inter-thèmes sur une carte thématique en s'appuyant sur un hypertexte généré automatiquement selon une technologie antérieure au World Wide Web.

Sur la base d'une telle carte thématique, deux types d'analyse de l'information sont considérés : l'une est l'observation de la structure du corpus de données et l'autre et l'observation du champ de recherche (qui fait quoi, où et quand ?).

Dans cette expérience, l'utilisation d'un hypertexte spécialisé dans la visualisation et l'exploration de cartes thématiques, illustre l'un des principes qui prévaudront à la conception du système HENOCH : l'utilisation de la carte comme moyen d'exploration des structures thématiques.

---

[1] Grivel L., Mutschke P., Polanco X. 'Thematic mapping on bibliographic databases by cluster analysis : a description of SDOC environment with SOLIS', Journal of Knowledge Organization, Vol. 22, n°2, 70-77, 1995
Cet article est le fruit d'une collaboration avec le 'InformationsZentrum Socialwissenschaften (IZS)' de Bonn. Il montre l'application d'une méthode d'analyse sur des données provenant d'une base allemande en sciences sociales (SOLIS) et illustre un mode d'exploitation des résultats que permet le système hypertexte

1. Introduction

Bibliographical information in public databases are, as Brookes (2,p.9) says, ”abundantly generated and systematically stored but not yet efficiently used”. The present paper addresses the problem of an end-user who is searching for information in a database. Usually, he needs to get an idea of the state of the art in his special domain of interest. In order to support the intellectual work of analysing retrieved documents in this respect, a coword-analysis method has been developed which discovers the thematical structure of a database and presents it as a map of themes on a graphical user interface. The SDOC-system from INIST (Institut de l´Information Scientifique et Technique) is an implementation of this method, and aims at mapping scientific research fields in large databases. Our goal is to demonstrate the thematic mapping facilities of SDOC with a German bibliographical database, here the SOLIS database of the Informationszentrum Sozialwissenschaften. SOLIS provides information mainly about German-language scientific literature, journal articles, contributions in compilations, monographs, and "grey literature".

Document-based retrieval systems normally use an indexing vocabulary to describe the content of its documents, and an online system to access these documents. The output of such a system in response to the user's query is a set of individual references. In this study, we imagine a French user who is searching for information in SOLIS concerning the field of social history in Germany. He selects all the literature processed over a three-year period (1989-90-91) in the SOLIS database having ”social history” as primary or secondary classification code and indexed by the keyword ”Germany”. This yields 285 bibliographical references. Traditionally, the user could only browse sequentially these documents with the difficulty of determining the importance of the topics and the links between them. By examining the indexing vocabulary, he can define certain topics manually and search for related documents. But even if the sample is not big, this iterative process is long and fastidious. The problem faced by all users of information systems is the need to reduce the amount of information to a manageable number of items to be examined.

SDOC belongs to a family of methods which use term associations and clustering techniques to solve this problem. Callon, Courtial, Turner and Bauin (3) call it "coword analysis" and Salton (12) "term clustering". This technique was early used in the SMART automatic document retrieval system (11). The use of term associations in automatic information retrieval has been studied since a long time, whereas coword analysis[2] has been implemented in the eighties into the LEXIMAPPE program to highlight the dynamics of scientific and technical development. In the latter context, cowords are used for identifying and visualizing the centres of interest in scientific literature by mean of coword maps (3).

[2] This method is an alternative to the well known tradition of citation analysis (9) and co-citation analysis (13); see (1) for a comparison of Co-Citation and Co-Word Clustering; see (7) and (4) for an introduction to scientometrics and scientific watch.

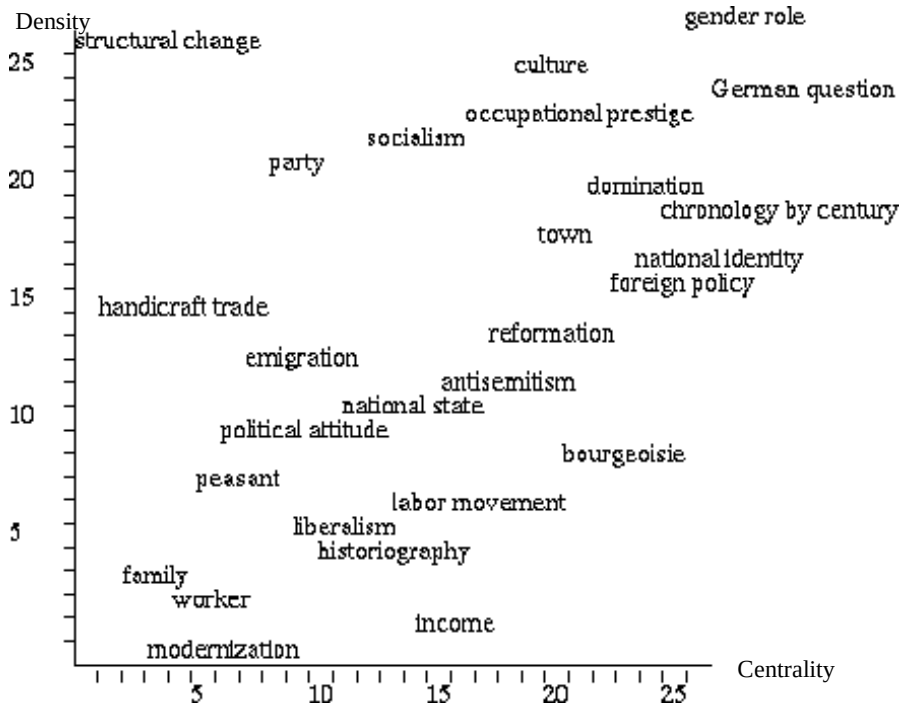


Figure 1-1: A General Map of Social History Themes

Like LEXIMAPPE, SDOC[3] produces a classification of *themes*, i.e. clusters of closely tied keywords, characterizing the domain studied, which can be the complete database or a subset of it referring to a special query. Such clusters are structured internally by means of relationships between the keywords of a cluster, and externally by interrelations between different clusters. The topics are visualized in a two-dimensional space or *Thematic Map* according to the semantic strength of their internal (*Density*: Y-axis) and external associations (*Centrality*: X-axis). Figure 1-1 shows an example of such a map of themes obtained from the 285 retrieved documents, saying, for instance, that *German Question*[4] was a central and intensively discussed theme of Social History research 1989-92. In this way, the user obtains an aggregation of thematic information. Furthermore, SDOC generates a hypertext system. Thus, the user can navigate through the generated knowledge space (map of themes). SDOC is described more detailed in Section 2.

On the basis of such thematical maps two types of *information analysis* can be considered: One is the analysis of the thematic structure of the database itself ("What is in the database?"), the other is the observation of the research field (”Who does what,

[3] SDOC differs from LEXIMAPPE concerning technical characteristics: SDOC has been implemented in C under UNIX, in order to allow the treatment of very large data files, whereas LEXIMAPPE is for DOS- and McIntosh-systems. The modules of SDOC rely on a library of C-functions, developed at INIST, specialized in the treatment of any SGML document (8), so that SGML is used by SDOC both as a conversion format for the raw data as input and as pivot format for the intermediary data which are exchanged between the modules.

[4] In the following, cluster names are printed in italics and start with an uppercase letter. Keywords are printed in italics, small letter size and lowercase letters.

where and when?”). A researcher or teacher in social history at least needs to know the thematic structure of the database he is consulting to satisfy his information request. The role of information analysis here is to provide the user with a state of the art of a certain domain of interest, in order, for instance, to get its most relevant scientists or journals, or to compare the scientific discussion in different countries (10).

In this paper, we will focus on the analysis of the thematic structure of the database. By applying SDOC to the SOLIS data file (see Section 3), we want to demonstrate how this tool can be used to support this kind of analysis on the basis of bibliographical data.

2. Thematic Mapping

2.1. Coword Analysis

Coword analysis used in SDOC is an analytical method for identifying and visualizing the centres of interest in scientific literature (3). The method is founded on the use of keywords as indicators of information content. The essential concept is the cooccurrence of content-describing keywords belonging to the same document. It is based on the idea that two keywords i and j which are used together in the description of a single document are related. It is clear that the cooccurrence value Cij (number of cooccurrences of words i and j in a given set of documents) is not the best measure of the strength of a keyword association because very frequently used keywords have an advantage over those used less often. In order to normalize the proximity value of keyword pairs the *Equivalence index* $Eij = Cij^2 / (Ci * Cj)$ (square of Ochiai index also called Salton index) is used, where Ci is the frequency of i and Cj the frequency of j in the data set. The keyword *German question*, for instance, cooccurs three times with the keyword *reunification*; thus, their association has an Equivalence index of 0.3, since *German question* has a frequency of ten, whereas *reunification* appears only three times in the datafile.

2.2 SDOC´s clustering process

These weighted coword-relations are the basis to construct a thematic representation (keyword clusters) of scientific areas and the relationships between research themes. The clustering-method aims at aggregating the keywords into groups of closely linked keywords. The algorithm implemented in SDOC is an adaptation of the single-link clustering in accordance with readability criteria: size of the cluster (minimum and maximum number of keywords belonging to it), and the maximum number of keyword associations constructing the cluster. The algorithm used is the following: Initially, each keyword is considered as a cluster. The list of keyword pairs, sorted by decreasing value of Equivalence index, is examined sequentially to build the clusters. If both elements of a given pair belong to the same cluster, the link between these keywords is considered as an internal association of that cluster. If they belong to two different clusters, the algorithm tries to aggregate the clusters into one by merging them. This is authorized if the size of the resulting cluster complies with the readability criteria. Otherwise, the association is taken to be an external association. Three saturation options are available when an aggregation fails because of the readability criteria: 1) forbid any new aggregation for these two clusters, 2) forbid any new aggregation of the larger of these two clusters, 3) do nothing.

The following example (see Figure 2-1) illustrates the building of the clusters *German Question* and *Foreign Policy* including their relationships (the links are valued by the Equivalence index of the respective keywords association). At a given time, *German Question* is composed of the links *Berlin <-> cold war*, *Berlin <-> reunification*, *cold war <-> german question*, *Berlin <-> german question*, *reunification <-> german question*, *german question <-> policy of detente*, *policy of detente <-> security policy*, *policy of detente <-> international relations*, *reunification <-> SED* and *GDR <-> SED*; the cluster *Foreign Policy* is only defined by *german policy <-> foreign policy*; and there is no link between these clusters. When the algorithm examines the associations *security policy <-> foreign policy* and *security policy <-> german policy*, the two clusters can not be merged because of the size criteria. Therefore, these links are stored as external associations. Each further association between keywords of *German Question* and *Foreign Policy*, such as *german question <-> german policy*, is represented as external link.

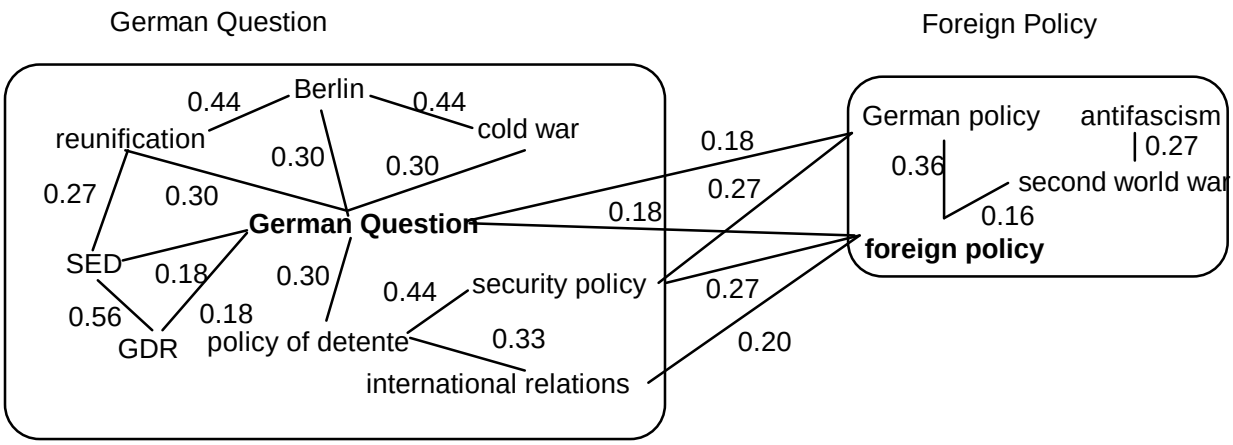


Figure 2-1: The building of clusters *German Question* and *Foreign Policy*

The user can modify the parameters used to compute the associations and construct the clusters. The goal here is to find a compromise between good readability of the results (not too many clusters) and what we accept to lose in terms of information. The parameters for this particular study are put in parenthesis.

<u>Indexing vocabulary :</u>
Minimum frequency of keywords (2)
Suppression of very frequent keywords (Germany)
<u>Associations :</u>
Selection of a statistical index (Equivalence index)
Minimum keyword cooccurrence (2)
<u>Clustering :</u>
Saturation strategy, i.e to saturate the largest cluster
Min. and max. size of clusters (4 and 10 keywords)
Max. number of internal and external associations (20)
Max.number of external associations (10)

2.3 The Structure of a Cluster

A cluster represents a special theme or centre of interest found in a set of documents. The keywords appearing in its internal associations are called *internal keywords*. The number of internal keywords defines the size of the cluster. Those keywords rejected during the clustering because they do not meet the "maximum cluster size" criteria are recorded as *external keywords*[5]. Each keyword has a weight indicating its centrality in the cluster. For a given cluster *C*, *N* being the number of internal and external associations and *Fi* the number of occurrences of term *i* in the associations, the weight *W(i)* of term *i* of cluster *C* is defined by *W(i) = Fi/N*. The internal keyword with the highest value is chosen to name the cluster automatically[6]. In the following the keywords defining the cluster *German Question* are shown:

Weight	Frequency	Keyword	
0.47	10	German question	
0.18	5	Socialist Unity Party of Germany	(SED)
0.18	3	security policy	
0.18	3	policy of detente	
0.18	3	reunification	
0.18	3	Berlin	
0.12	9	German Democratic Republic	(GDR)
0.12	4	international relations	
0.12	3	cold war	
0.18	5	foreign policy*	
0.12	5	Germany policy*	

The Equivalence indices of the *internal associations* describe the strength of the keyword associations defining the internal structure of a cluster. In order to have an indicator of its degree of cohesiveness (*Density*), the mean value of the internal associations is used (density of *German Question*: 0.34). The *external associations* are the associations existing between the keywords of this cluster (internal keywords) and keywords belonging to other clusters (external keywords). The mean value of the external associations of a cluster (*Centrality*) is an indicator of its degree of dependance with regard to other clusters (centrality of *German Question*: 0.22). The *saturation threshold* of a cluster is the Equivalence index of the last internal association added before the cluster becomes saturated (the saturation threshold of *German Question* is 0.27). This value characterizes the relationship between density and centrality of a theme. The centrality index of *German Question*, for instance, is below its saturation threshold, showing that this theme can be extended to *Foreign Policy*. The saturation threshold is therefore an important information for interpreting interrelations between clusters (see Section 3.4 Analysing Cluster Relationships).

The number of external associations displayed for a given cluster may be limited. This is one parameter of the application. Thus, the external associations are not necessarily bidirectional. We introduce the idea of *thematic reference* to indicate the number of

[5] indicated by a star in the example.

[6] This is only a label suggested by our program. It may be changed if it is not felt to be appropriate to the cluster.

times that keywords of one cluster appear in the external associations of other clusters. When a cluster refers to another one by its external associations, the latter is said to be *referenced* by the former as a related item of information. Here, *German Question* is referenced 13 times by other clusters indicating that its influence goes beyond the topic described by the keywords of the cluster (Section 3.3 illustrates these relationships).

Considered as a classification unit, a cluster gathers together not only keywords, but also a set of documents. A document is assigned to a cluster if it is indexed by a couple of two internal keywords or a couple of one internal and one external keyword of the cluster. A document may therefore belong to several clusters. A relevance weight is computed for each document. This is the sum of the weights of keywords in the cluster indexing the document, divided by the number of keywords belonging to it. In the following, the documents dealing with the *German Question* topic are shown:

Weight	Title:
0.14	The social-democratic intra-party discussion on security, detente and German unity
0.11	Between the Cold War and detente : security and Germany policy within the system of the allied powers in the years 1953-1956
0.11	From "civil war" to the responsible community
0.10	The four-sector city of Berlin in the German press 1945-1949
0.10	Attitude of the SED and the GDR towards German unity 1949-1987
0.08	The German policy of the government of the U.S.A. in preparation and during the course of the Potsdam Conference
0.07	The Socialist Unity Party of Germany (SED) and the national issue
0.07	Neither a hammer nor an anvil? : observations on the present-day situation in Germany (1973)
0.06	Contributions on the history of the Berlin democracy : 1919-1933/1945-1985
0.05	The Socialist Unity Party of Germany (SED) in history and the present age
0.05	The political obstruction to modernization in France during the interwar period
0.03	On the appearance of the first volume of the "History of the SED"
0.02	The German-Japanese relations during the Third Reich
0.02	The Socialist Unity Party of Germany (SED) and German history

Additional information such as a list of authors, a list of sources (journals, books etc.) or institutional affiliations, can also be assigned to the clusters if this information is in the bibliographical reference. The weight assigned to each item is the sum of the weights of the documents where the item appears.

2.4 Constructing Thematic Maps

The measures of Density and Centrality allow the visualization of themes and their relationships in a two-dimensional space (map), where the x-axis corresponds to Centrality and the y-axis to Density[7]. In order to support a consultation of the clustering results, SDOC integrates this map in a graphical hypertext-based user interface (s. Fig. 2-2).

[7] To avoid recovering clusters having similar coordinates on the map, the software also makes it possible to plot the clusters by rank along these two axes.

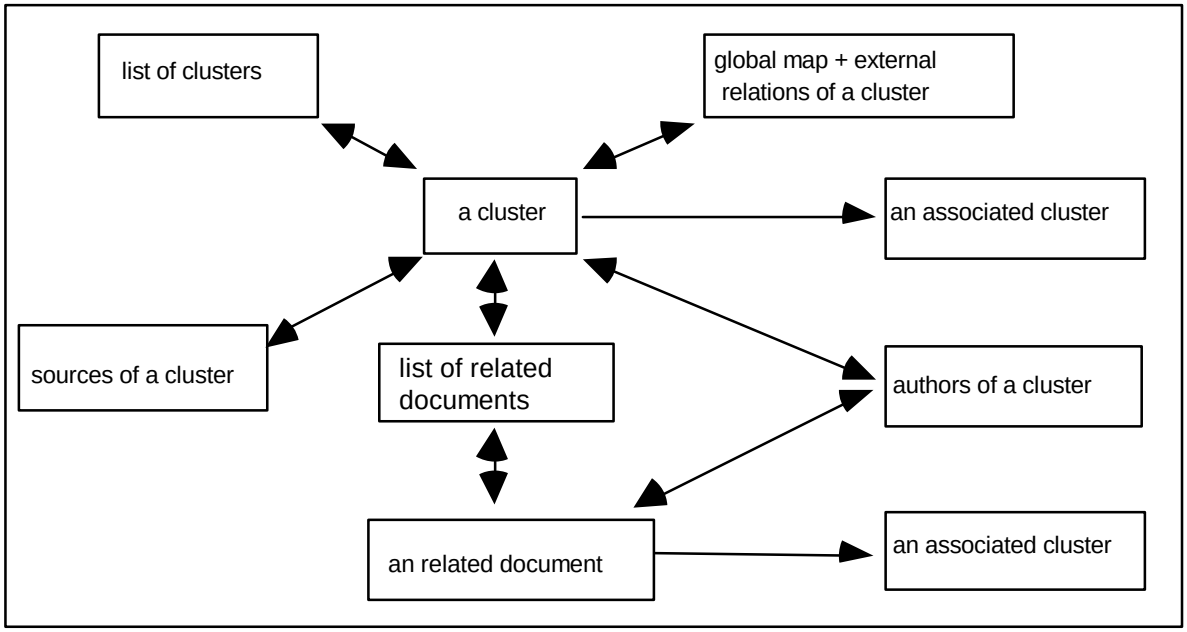


Figure 2-2: Browsing the organization of a topic,
the key figures and the sources of information

The starting point for the navigation is the list of clusters sorted by saturation threshold. This corresponds to the order in which they have been ”frozen” during the clustering. The user selects the cluster name and points to its description. He can then examine: a) the characteristics of the cluster (number of documents, authors and sources, saturation threshold, density, centrality, number of citations by the other clusters); b) the characteristics of the keywords in the cluster (weight, frequency) and their associations (Equivalence index, cooccurrence); and c) the associated clusters including a description of the external associations involved.

3. Information Analysis of the SOLIS Datafile

3.1 The Indexing Vocabulary

Keywords are primarily used for information retrieval by boolean queries. Here, they are used as content indicators to which the SDOC analysis is applied. The vocabulary indexing the 285 retrieved Social History documents consists of 892 controled terms manually assigned on the base of the Social Science thesaurus of the Informations-zentrum Sozialwissenschaften. For this coword analysis, the English keywords of SOLIS are used, with the exception of the keyword ”Germany”, because, given the search query, this keyword yields no information. The 499 keywords of frequency 1, which represent 56 % of the indexing vocabulary, are excluded as input to the coword analysis. They complicate the keywords association network with potentially noisy information. So the effective number of keywords as input to the clustering is 392.

In order to analyse this datafile, we will first study the variables which characterize a cluster as an indicator of a research theme. Then we will focus on the use of the hypertext maps as a means to explore the thematic structure of the database by theme.

Finally, we will analyse the cluster relationships.

3.2 Coword Clusters as Knowledge Indicators

Applying SDOC on the Social History document set provides 27 clusters in all (s. Fig. 1-1: A General Map of Social History Themes). Table 3-1 shows these clusters with the following characteristic data:

[1] Cluster saturation threshold
[2] Density
[3] Centrality
[4] Number of internal keywords
[5] Number of external keywords
[6] Number of internal associations
[7] Number of external associations with other clusters
[8] Number of thematic references of a subject by other topics
[9] number of bibliographical references related to the cluster
[10] number of bibliographical references exclusively related to the cluster

Table 3-1: Characteristics of the 27 clusters obtained (in alphabetical order)

Name	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Antisemitism	0.125	0.212	0.106	10	4	16	4	4	22	3
Bourgeoisie	0.133	0.185	0.129	9	6	12	7	24	89	1
Chronology by Century	0.200	0.296	0.160	9	7	13	7	38	188	21
Culture	0.173	0.376	0.122	10	5	14	6	5	19	0
Domination	0.118	0.296	0.131	8	5	10	8	10	25	1
Emigration	0.083	0.218	0.071	10	1	18	1	6	22	6
Family	0.111	0.148	0.033	4	9	3	10	1	12	1
Foreign Policy	0.160	0.262	0.143	4	7	3	10	6	11	1
Gender Role	0.213	0.527	0.196	8	2	18	2	5	10	2
German Question	0.267	0.337	0.219	9	2	12	5	13	14	0
Handicraft Trade	0.167	0.222	0.019	5	4	4	10	0	12	1
Historiography	0.082	0.163	0.086	8	8	7	9	2	18	2
Income	0.114	0.137	0.103	9	5	13	7	6	19	0
Labor Movement	0.091	0.169	0.096	9	8	10	10	7	46	1
Liberalism	0.062	0.166	0.079	7	6	6	9	2	18	1
Modernization	0.071	0.093	0.039	4	6	3	9	0	16	2
National Identity	0.188	0.289	0.147	9	2	14	2	8	19	6
National State	0.078	0.194	0.087	9	10	10	10	9	33	0
Occupational Prestige	0.190	0.315	0.115	9	6	12	6	8	18	0
Party	0.133	0.297	0.076	6	7	11	9	3	11	2
Peasant	0.089	0.184	0.060	7	7	9	10	0	14	0
Political Attitude	0.114	0.186	0.066	5	6	4	9	2	13	0
Reformation	0.111	0.221	0.121	8	3	14	6	4	13	0

Socialism	0.167	0.309	0.095	8	5	8	10	6	15	0
Structural Change	0.200	0.486	0.000	8	0	20	0	5	4	0
Town	0.113	0.289	0.124	10	6	12	7	16	60	1
Worker	0.067	0.142	0.057	6	8	6	8	1	15	0

Column [1] permits to identify the order in which the clusters have been ”frozen” during the clustering. It is used in combination with column [3] for analysing cluster relationships (see Section 3.4). The values of columns [2] and [3] are used to plot the clusters in a two-dimensional space representation. To get a more detailed idea of the structural diversity of the clusters, a connection can be made between these mean values [2] and [3], and the number of internal and external associations [6] and [7] of each cluster.

The cluster size [4] is the number of distinct keywords appearing in the internal associations [6] whose mean value [2] represents the density of the cluster. This characterizes the cohesion of the cluster. The sum of the values of column [4] gives the number of keywords kept in the clusters. Here 208 keywords appear in the 27 clusters. This can be compared with the initial number of keywords (892) to evaluate the "data reduction".

The number of external associations [7], the mean value of these associations [3], the number of external keywords involved in these external associations [5], and the number of times a cluster is referenced by the others [8] give an idea regarding the role it plays within the network of themes describing a certain research context (see Section 3.4 Analysing Cluster Relationships).

Column [9] and [10] indicate the quantity of bibliographic information relative to each cluster. Since document classes can overlap, the total number of documents classified in a given cluster [9] is not the same as the number of documents exlusively associated to that cluster [10]. The sum of the values of [9] gives the number of documents belonging to the clusters. In this case, there are 756 document cluster associations, whereas the total number of distinct documents in the clusters is only 266. Of these 266, 52 are related to exclusively one cluster. Overlaps like this are indicators of theme relationships. More than 93% of the documents in the initial file of 285 documents are covered by the 27 clusters. We may stress that we have obtained a manageable number of items (27 clusters) without losing too much bibliographic information.

3.3 Mapping Knowledge: A Hypertext System

On our maps (s. Fig. 3-1 to 3-4), the 27 clusters are arranged along the vertical Y-axis by order of increasing mean value of internal associations (density), and along the horizontal X-axis by order of increasing mean value of the external associations (centrality). Each cluster has a certain thematic significance within the studied research field expressed by its position on the two axes. The fact that two clusters appear close to one another in the information space (or map) does not mean that they are closely associated with one another. It only means that their values of centrality and density are similar.

The higher a cluster is located on the Y-axis, the more it is a coherent unit of

information. The farther right it is on the X-axis, the greater are its links to other clusters. The authors of coword analysis method traditionally distinguishes four types of clusters: clusters with high density and centrality (type 1), with a low density and high centrality (type 2), with high density while peripheral from the point of view of centrality (type 3), and themes with low values on both axes (type 4). Callon, Courtial, Turner and Bauin (3) call this representation "strategic diagram" and use this typology to assess the strategic interest of the themes. In this kind of analysis, the mainstream themes in the research field studied should be represented by those clusters having the highest values on both axes (type 1 in table 3-2). Clusters of type 2 may correspond to central themes in the future. Clusters of type 3 are specialized themes while clusters of type 4 are both peripheral and weakly developed and represent the margins of the network. This categorization should be cautiously used in collaboration with an expert of the domain. The strategic diagrams are generally used to study the life cycle of the themes. A case study can be found in (6).

Here, our use of the map is different. We use this representation to define an informational space or global context of research information where the local networks are highlighted, i.e. the associations between the clusters. The hypertext interface permits the user to follow the local networks of each theme (s. Fig. 3-1 to 3-4), and then to proceed to an analysis. If, for instance, he is interested in questions of nation and nationality in the framework of the *German Question*, he can see that this cluster (s. Fig. 3-1) is associated with one other cluster, *Foreign Policy*.

Type 1	Gender Role, Culture, German Question, Occupational Prestige, Domination, Sixteenth Century, Town, National Identity, Foreign Policy, Reformation
Type 2	Antisemistism, Bourgeoisie, Labor Movement, Income
Type 3	Structural Change, Socialism, Party, Handicraft Trade
Type 4	Emigration, National State, Political Attitude, Peasant, Liberalism, Historiography,Family, Worker, Modernization

Table 3-2: Cluster categorization in a strategic diagram

German Question and *Foreign Policy* are associated by way of five bidirectional associations (s. Fig. 2-1). The analysis of these associations shows that *Foreign Policy* is a subtheme of *German Question* because the saturation threshold of *German Question* is higher than the mean value of its external associations to *Foreign Policy*, and, vice versa, the strength of the external associations of *Foreign Policy* with *German Question* are higher than its saturation threshold. The relative position of *Foreign Policy* with respect to *German Question* (below, and more left) is an indicator but not a sufficient condition for the existence of such a relationship, because we need to know the saturation threshold and the strength of the external associations concerned. Figure 3-2 illustrates the local network of the theme *Foreign Policy*. Thus, the initial topic *German Question* is also associated with *National Identity*, *Labor Movement* and *Emigration*.

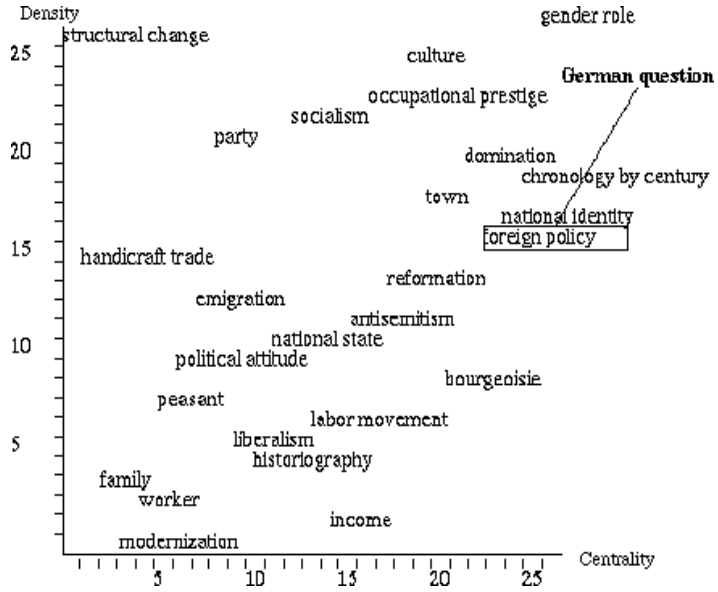


Figure 3-1: Cluster *German Question*

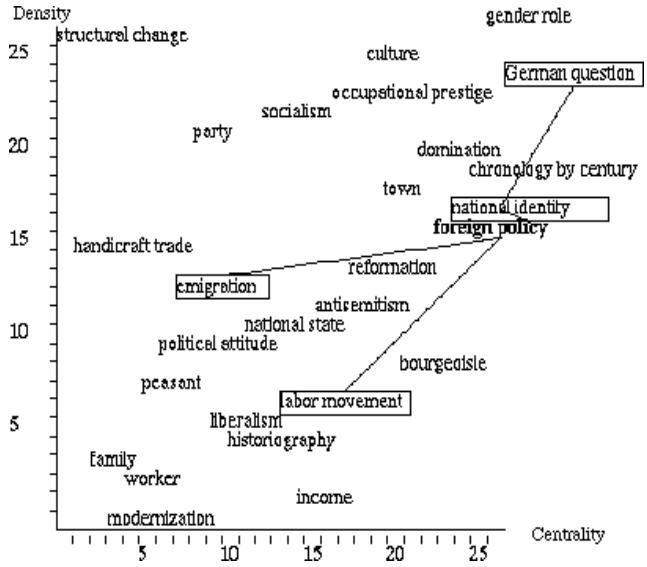


Figure 3-2: Cluster *Foreign Policy*

Suppose the user is now interested in the position of *National Identity*. Figure 3-3 shows that this topic is associated with the initial theme *German Question*, and refers to a new topic, *Socialism*. *National Identity* contains the keywords: *national identity, national consciousness, historical awareness, conception of history, German, Nazism, Hitler, Third Reich, nationalism*. It has external associations with *German Question*, by *conception of history - Socialist Unity Party of Germany (SED)*, and with *Socialism*, by *Nazism - socialist party*.

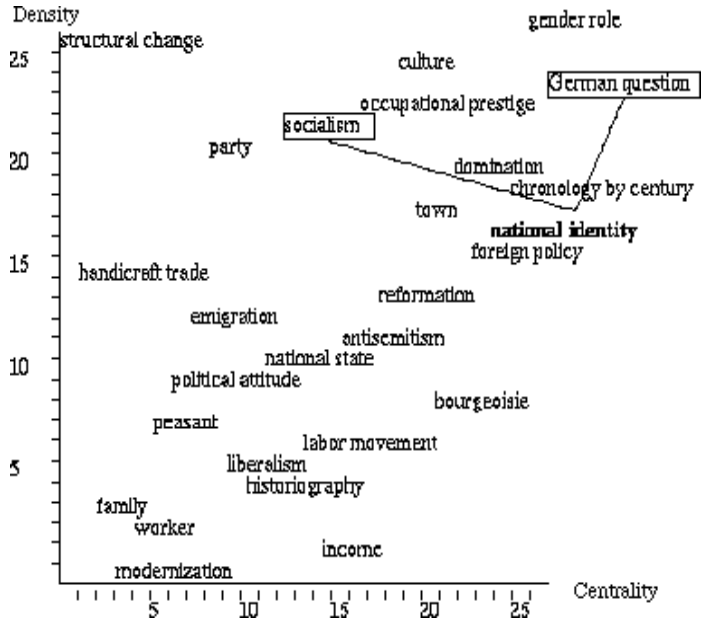


Figure 3-3: Cluster *National Identity*

The *Socialism* cluster refers back to *Labor Movement* and *National Identity*, and opens the network towards two other themes, *Party* and *Chronology by century* . Moving from one topic to another, the user explores the content of his data by examining a structured knowledge space. He can decide either to follow another informational network or to stop the navigation process and browse the literature aggregated under a topic.

3.4 Analysing Cluster Relationships

Coword analysis is not only a method for classifying bibliographical references in clusters representing a research theme. It also provides the possibility of analysing the associations between themes. This analysis relies on the distinction between internal and external associations, the notion of cluster saturation threshold, and the size of the clusters.

Table 3-3 describes two categories of clusters:
[A] those whose external associations mean value is higher than the saturation threshold, i.e. the external links are as strong as the most internal associations;
[B] those whose external associations mean value falls below the saturation threshold, i.e. the internal links are much stronger than the external associations. In this latter category, we distinguish between those whose external associations are, nethertheless, relatively strong [B1] from those whose external links are very weak [B2].

Clusters of category [A] identify themes which are secondary (in the datafile) insofar as they are of weak internal cohesiveness, whereas their associations with other clusters are relatively strong, i.e. they seem to be subthemes of these clusters. For instance, *Liberalism* seems to be secondary with respect to the theme *Bourgeoisie*. Furthermore, in this category of clusters, we can discover crossroad clusters (*Domination* and *Town*) which connect very heterogenous topics via one generic keyword (s. Fig. 3-4). Thus, crossroad clusters usually represent very generic research topics, which are crossing

points of themes.

Clusters of category [B1] could be qualified as mainstream themes if their internal associations are numerous and relatively strong. A typical example is *German Question* (s. Fig. 3-1 and 3-2) whose local network has been already studied. An analysis process should start with them because they are the main thematic nodes of the network.

Clusters of category [B2] represent peripheral themes because the links tying them to the network are very weak. In this category, *Handicraft Trade* is a good example of such a cluster. The only external associations it has are with *Chronology by century*. *Family*, *Worker* and *Modernization* have numerous but weak associations to other clusters. Since their internal structure is, moreover, very weak (see the number of internal keywords [6] and internal associations [7] in table 3-1), we consider them as peripheral themes. *Structural Change* is a special case, because it points out a theme with a strong density, i.e. a homogeneous research field, but without any association with other clusters.

SDOC visualizes such thematical networks in the form of maps. In other words, it maps the knowledge embedded in documents (thematic structure), but also the individual agents (authors, institutions) and the way they communicate. By considering the relationships between clusters, their internal structure and the less or more central role they play within a network of themes the importance of a certain thematic aspect for the research field studied can be examined.

Table 3-3: Categories of clusters

A	Domination, Town, Reformation, National State, Labor Movement, Liberalism, Historiography
B1	Gender Role, Culture, German Question, Occupational Prestige, Socialism, Party, Sixteenth Century, National Identity, Foreign Policy, Emigration, Antisemitism, Political Attitude, Bourgeoisie, Peasant, Income
B2	Structural Change, Handicraft Trade, Family, Worker, Modernization

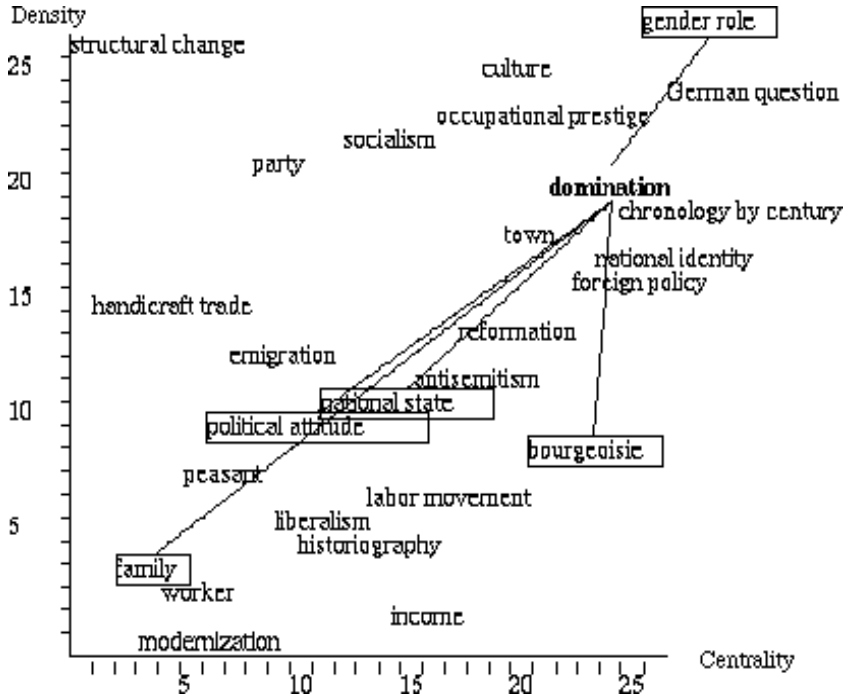


Figure 3-4: Cluster *Domination*: An example of crossroad cluster

4. Conclusion

In the present paper, two possibilities of using the mapping method of SDOC are illustrated. The first one is to give an easy access to distributed database information. In front of the thematic structure of the database content the user can define his own strategy of information search for the problem he has to solve. He may discover relations between themes he would not have thought of; and on this basis he can adjust his query. The second method is to use such *Thematic Maps* as a means of analysing information. Besides the traditional way of analysing a coword map as a strategic diagram, which reflects only two parameters characterizing the clusters (centrality and density), we have introduced the clusters relationships analysis taking into account further important parameters of the clustering: the saturation threshold, the size of the clusters, and the number of associations. Since this approach avoids some interpretation problems due to the criteria of cluster size, it provides a more adequate interpretation of links between themes.

Our objective was to implement an environment which offers the user a contextual view of the informational space contained in a set of bibliographical references, so that he can locate his demand of information more precisely. Since we are working at a level of indicators, we are not concerned with exactness. A specialist in the field will always have the final say concerning the results of an automatic information analysis. Our intention is to provide him with a working tool to support his own information discovering process, with the possibility of going beyond his special subject in order to explore neighbouring domains. We believe that such an environment best arms the user to face the growing volume of information.

Acknowledgments: We are grateful to our INIST and IZ colleagues, and particularly to M. Herfurth (head of the IZ research department), for their valuable comments.

5 References:

(1) R.R. Braam, H.F. Moed, A.F.J. van Raan : ”Comparison and Combination of Co-Citation and Co-Word Clustering”, in: Select Proceeding of the First International Workshop on Science and Technology Indicators, Leiden, 14-16 November 1988, p. 307-337.
(2) B.C. Brookes : ”The foundations of information science. Part. IV. Information science: The changing paradigm”, Journal of Information Science 3 (1981), p. 3-12.
(3) M. Callon, J-P.Courtial , W.A.Turner , S.Bauin : ”From translation to problematic networks: an introduction to co-word analysis”, Social Science Information 22 (1983), pp. 191-235.
(4) M. Callon, J-P.Courtial , H. Penan : ”La scientométrie”, Presses Universitaires de France, collection ”Que sais-je”, Paris, 1993.
(5) M. Callon, J. Law, A. Rip (eds).: Mapping the dynamics of science and technology, London: The Macmillan Press Ltd, 1986.
(6) M. Callon, J-P.Courtial , F. Laville : ”Co-word Analysis as a tool for describing the network of interactions between basics and technological Research: the case of polymer chemistry”, Scientometrics 22 (1991), No1, pp. 155-206.
(7) H. Desvals, H. Dou : "La veille technologique", DUNOD, Paris 1992.
(8) J. Ducloy, P. Charpentier, C. Francois, L. Grivel : ”Une boite à outils pour le traitement de l’information scientifique et technique”, Génie logiciel et systèmes experts 25 (1991), pp 80-90, Paris.
(9) E. Garfield : "Citation analysis as a tool in journal evaluation", Science 178 (1972), pp 471-479.
(10) X. Polanco X., L.Grivel : ”Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United State of America), 4th International conference of Bibliometrics, Informetrics and Scientometrics - 11-15 Septembre 1993, Berlin, Germany.
(11) G. Salton : "The SMART retrieval system - Experiments in automatic document processing", Englewoods Cliff, New Jersey, Prentice Hall Inc., 1971.
(12) G. Salton : "Automatic text processing : the transformation, analysis and retrieval of information by computer, New York, Addison Wesley, 1989.
(13) H. Small, "Co-citation in the scientific litterature: A new measure of the relationship between two documents", Journal of the American Society of Information Science 24 (1973), pp. 265-269.

Chapitre 5[1]

Démarche générale d'application de méthodes d'analyse de l'IST et d'exploitation de leurs résultats

Comme cela a été signalé dans le chapitre 2, l'analyse de l'IST ne peut être effectuée sans s'appuyer sur une solide démarche méthodologique. Ceci suppose une documentation adéquate de la ou des méthode(s) employée(s) et de la chaîne de traitement, une définition claire des sources de données et des indicateurs utilisés. C'est l'approche qui est suivie dans ce chapitre.

Ce chapitre décrit une démarche d'analyse mettant en œuvre deux méthodes permettant de classer et représenter graphiquement d'énormes quantités d'information bibliographique: les mots associés, et une autre plus récente associant une technique de classification, les K-means axiales, à une technique d'analyse factorielle courante : l'Analyse en Composantes Principales (ACP).

Dans la première partie, les deux méthodes sont présentées en détail et comparées d'un point de vue théorique et pratique. Bien qu'il existe une grande symétrie entre les deux processus, expliquant les accords observés expérimentalement entre les résultats des deux méthodes, les méthodes offrent des représentations différentes : classes de mots-clés structurées par les relations de cooccurrences dans un cas, classes de mots-clés floues et recouvrantes dans l'autre; cartes thématiques fournissant des informations de natures différentes : indicateurs structurels et visualisation des réseaux locaux dans un cas, oppositions des thèmes selon deux axes principaux dans l'autre cas.

La deuxième partie de ce chapitre aborde le problème de la qualification des résultats afin de limiter les risques d'erreurs lors de leur interprétation. Une démarche d'analyse est proposée qui met l'accent sur les apports de la navigation hypertexte et sur la possibilité de mesurer les accords entre les résultats des deux méthodes d'analyse par des indicateurs globaux.

Néanmoins, comme les souligne la conclusion de ce chapitre, les hyper-documents générés automatiquement restent statiques, ce qui ne permet pas de croiser dynamiquement certaines informations relatives aux résultats de classification et aux données à analyser. L'idée vient alors de constituer une base de données accessible via le Web où sont stockés tous les éléments nécessaires à l'analyse de l'information. C'est le concept de base de données infométriques qui est développé dans le chapitre suivant.

---

[1] Grivel L., Francois C. 'Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique' - Solaris *n°2* "Les sciences de l'Information : Bibliométrie, Scientométrie, Infométrie", Presses universitaires de Rennes, p.81-113, 1995 (http://www.info.unicaen.fr/bnum/jelec/Solaris).
Cet article a été publié dans la revue électronique SOLARIS éditée par le Groupe interuniversitaire de recherche en sciences de l'information et de la communication (GIRSIC) dans le cadre d'un numéro spécial sur la bibliométrie, scientométrie, infométrie. Visant à mettre en perspective des approches infométriques, ce numéro a rassemblé un ensemble de réflexions pratiques (Grivel L., Francois C., Lelu A.) et théoriques (Polanco X.) dans le développement des systèmes d'information (Barre R., Laville F. Teixera N., Zitt M.), de nouveaux modes d'écritures (Noyer J.M., Courtial J.P), la production de connaissances (Turner W., Bossy M.).

1- Introduction

Dans un contexte de veille scientifique, l'analyse infométrique de l'information scientifique et technique comprend non seulement une analyse de contenu à partir des mots-clés, résumés et titres mais aussi une analyse de ses acteurs, leurs relations, leurs moyens de communications (revues, rapports, congrès, ...), son actualité. Dans cette perspective, nous présentons ici une station d'analyse de l'information scientifique et technique développée dans le cadre du programme de recherche en infométrie de l'INIST/CNRS. D'un point de vue fonctionnel, elle doit non seulement fournir tous les indicateurs numériques usuellement mis en oeuvre pour prendre la mesure de l'information bibliographique, mais également proposer des représentations du contenu de la production scientifique. Elle automatise l'élaboration des distributions bibliométriques (statistiques unidimensionnelles sur les champs bibliographiques), et supporte deux méthodes permettant de construire des cartes thématiques : une méthode éprouvée, les mots associés [CALLON *et al* 1983]), et une autre plus récente associant une technique de classification, les K-means axiales [LELU 1990 et 1993] à une technique d'analyse factorielle courante : l'Analyse en Composantes Principales (ACP). Notre objectif est de classer et représenter d'énormes quantités d'information bibliographique afin d'en extraire des synthèses élaborées utilisables pour effectuer une veille scientifique (données chiffrées caractérisant un ensemble de références bibliographiques, hypertextes thématiques, documents de synthèse tels que des cartes de l'information scientifique et technique).

La première partie de cet article décrit les méthodes mises en oeuvre pour représenter le contenu de l'information et montre leur spécificité et leur complémentarité. Nous y exposons également nos choix technologiques, puis nous décrivons l'objet technique réalisé : une chaîne de traitement infométrique sous Unix, basée sur la norme SGML.

La deuxième partie est consacrée à l'analyse des résultats. Nous abordons ici le problème de la qualification des résultats afin de limiter les risques d'erreurs lors de leur interprétation. L'analyse des distributions bibliométriques n'est qu'esquissée. Elle ne présente, à notre avis, pas de difficultés majeures, puisqu'il est possible de s'appuyer sur des lois qui décrivent leur comportement. Par contre, l'exploitation des résultats de méthodes d'analyse de données demande quelques précautions car il ne faut pas oublier qu'elles procèdent par réduction de données. Nous exposons donc une démarche d'analyse basée sur l'observation d'indicateurs permettant d'apprécier la qualité des résultats produits par notre station de travail. Pour illustrer cette démarche, nous utilisons les résultats du traitement d'un petit corpus[2] de références bibliographiques (quelques centaines de documents).

En conclusion, nous effectuons un bilan comparatif des deux méthodes et décrivons les évolutions futures de la station de travail.

2 - Choix méthodologiques et technologiques

2.1 - Méthodes mises en oeuvre

[2] Il est entendu que nous l'appliquons également pour le traitement de gros corpus.

Si les méthodes à mettre en oeuvre pour obtenir les distributions bibliométriques sont relativement bien standardisées et banalisées [POLANCO 95], il n'en est pas de même pour la représentation de l'IST. C'est pourquoi nous nous contenterons de développer ce deuxième aspect.

Les indicateurs que nous utilisons pour représenter le contenu de l'information sont les cartes thématiques. D'une manière générale, nous définissons une carte thématique comme étant une représentation de la topologie des relations entre des disciplines ou des thèmes de recherche, telle qu'elles sont matérialisées sous la forme de données bibliographiques. Pour construire ces cartes, notre choix s'est porté en priorité sur deux méthodes d'analyse de corpus documentaire déjà décrites dans la littérature : la méthode des mots associés implémentée par le logiciel SDOC et une méthode associant les K-means axiales à une Analyse en Composantes Principales (ACP) implémentée par le logiciel NEURODOC.

Pour des raisons historiques, ces méthodes sont bien connues de notre programme de recherche. Nous bénéficions de l'expérience acquise par le SERPIA[3] , département de R & D du CDST[4] avant la fondation de l'INIST. En effet, la méthode des mots associés est le fruit d'une collaboration entre le Centre de Sociologie de l'Innovation de l'Ecole des Mines de Paris et le CDST [CALLON *et al* 1983]. Le logiciel développé à l'époque s'appelle LEXIMAPPE. Quant à la méthode basée sur les K-means axiales et l'ACP, elle a été mise au point par A. LELU, alors qu'il était membre du SERPIA [LELU 1990].

Ces deux méthodes utilisent les mots-clés qui indexent les références bibliographiques pour construire les structures thématiques "enfouies" dans les bases de données. Pour schématiser, elles trouvent les thèmes abordés et classent les documents selon ces thèmes. Ceux-ci sont ensuite disposés sur un espace à 2 dimensions : "carte thématique".

Les mots associés [CALLON et al. 1983, 1986, 1993] [COURTIAL 1990]

Cette méthode considère les mots-clés comme des indicateurs de connaissance (contenu des documents indexés) et se base sur leur cooccurrences pour mettre en évidence la structure de leurs relations (clusters[5]). L'idée de cooccurrence est essentielle. En effet, si on considère que deux documents sont proches parce qu'ils sont indexés par des mots-clés similaires, alors deux mots-clés figurant ensemble dans un grand nombre de documents seront considérés comme proches. Cependant, la cooccurrence ne permet pas à elle seule de mesurer la force des associations entre mots-clés (leur proximité), car elle avantage les mots-clés de haute fréquence par rapport à ceux de basse fréquence. L'emploi d'un indice statistique approprié permet de normaliser la mesure de l'association entre deux mots-clés. En pratique, nous utilisons le plus souvent l'indice d'Equivalence dont les valeurs varient entre 0 et 1: $E_{ij} = C_{ij}^2 / (C_i * C_j)$; où C_{ij} est le nombre de cooccurrences des mots-clés i et j, C_i la fréquence du mot-clé i, C_j la fréquence du mot-clé j.

[3] SERPIA : Service d'Etude et de Réalisation de Produits d'Information Avancés.
[4] CDST : Centre de Documentation Scientifique et Technique du CNRS.
[5] Un cluster est une classe de mots entre lesquels il existe des associations fortes.

A partir des mesures de proximité entre les mots, un algorithme de **classification hiérarchique** construit des groupes de mots proches les uns des autres (clusters) n'excédant pas une taille maximale (nombre de mots) fixée par l'utilisateur. Ainsi la figure 1 montre deux clusters C1 et C2 contenant respectivement les mots-clés A, B, C, D, E d'une part F, G, H, I d'autre part. Un cluster est donc constitué de mots associés les uns aux autres (associations internes). Les clusters peuvent avoir des relations entre eux. Ceci se produit lorsqu'il existe une association entre 2 mots-clés appartenant à 2 clusters différents (association externe) et que la taille du nouveau cluster qui aurait résulté de la réunion de ces 2 clusters dépasse la taille maximum définie par l'utilisateur. Ainsi C1 et C2 sont reliés par une association externe entre C et F car la taille des clusters ne peut excéder un maximum de cinq mots dans l'exemple présenté.

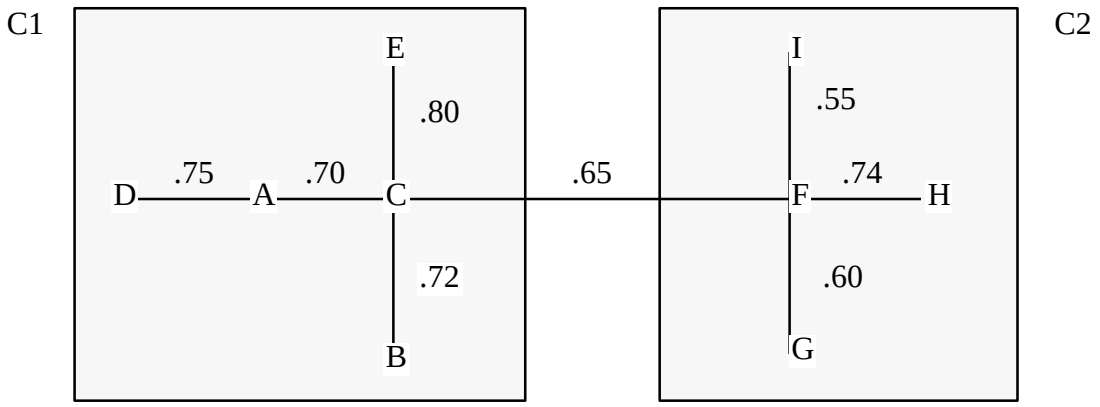


Figure 1 : deux clusters C1 et C2 de 5 mots maximum

Après le processus de classification des mots-clés, les documents sont affectés aux clusters de la manière suivante : un document est associé à un cluster, si dans sa liste de mots-clés, il existe au moins un couple de mots-clés qui pourrait constituer une association interne ou externe du cluster.

La classification est principalement paramétrée par le nombre maximal de mots pouvant constituer un cluster. C'est une variante de la procédure statistique habituelle qui consisterait à utiliser un seuil fixe (une "distance limite" à partir de laquelle aucune agrégation n'est plus effectuée). C'est un moyen pratique pour moduler la coupure dans l'arbre de classification (dendrogramme). En conséquence du critère de taille maximale, les classes résultantes sont très hétérogènes en densité. La première classe obtenue sera constituée des mots-clés les plus fortement liés alors que la dernière sera très lâche, restituant en cela la structure du réseau d'associations. On peut également limiter le nombre d'associations intra ou inter-clusters dans un souci de lisibilité. Les autres paramètres de la méthode se situent en amont de la classification (filtrages au niveau du vocabulaire d'indexation : fréquence des mots-clés, cooccurrence, ...), ou en aval (filtrage des clusters par le nombre de mots ou de documents qu'ils comportent, ...).

Cartographie

Des indicateurs structurels sont ensuite calculés. Ce sont la densité (valeur moyenne des associations entre mots-clés formant un cluster ou associations internes) et la centralité (valeur moyenne des associations entre les mots qui le constituent et les mots d'autres clusters ou associations externes). Ces valeurs sont ensuite utilisées pour positionner les clusters sur une carte. On peut ainsi repérer les thèmes (ou clusters) les mieux structurés du point de vue de leur densité (ou cohésion), les mieux rattachés au réseau (centralité). Sur une telle carte, la proximité entre deux thèmes indique qu'il sont structurellement proches, mais leur contenu sémantique ne sont généralement pas voisins. Les auteurs de la méthode des mots associés appellent ce type de carte "diagramme stratégique" [CALLON et al.1993, p86]. Ils l'utilisent pour évaluer l'intéret stratégique des thèmes. Leur objectif est avant tout sociologique : étude des dimensions sociales et organisationnelles de la science [COURTIAL 90], [TURNER 94]. Nous utilisons la même méthode de construction de cartes avec un autre objectif : permettre à un utilisateur d'appréhender globalement et localement le contenu d'un corpus bibliographique. Ainsi la figure 6 présentée dans la deuxième partie est un exemple de carte affichant les relations qu'un thème entretient avec d'autres thèmes, dans le domaine des systèmes experts et intelligence artificielle.

La méthode basée sur les K-means axiales et l'ACP [LELU 1990, 1993]

Cette méthode considère l'ensemble des références bibliographiques comme un nuage de points plongé dans un espace géométrique où chaque dimension correspond à un mot-clé. Elle est caractérisée par une représentation des classes par des vecteurs pointant vers les zones de forte densité du nuage.
La figure 2 montre l'exemple d'un corpus de documents indexés par les 3 mots-clés x_1, x_2, et x_3. Ces mots-clés définissent l'espace $\mathbf{R}^3$, un document **i** indexé par les mots-clés x_1 et x_2 aura les coordonnées suivantes : (1, 1, 0)

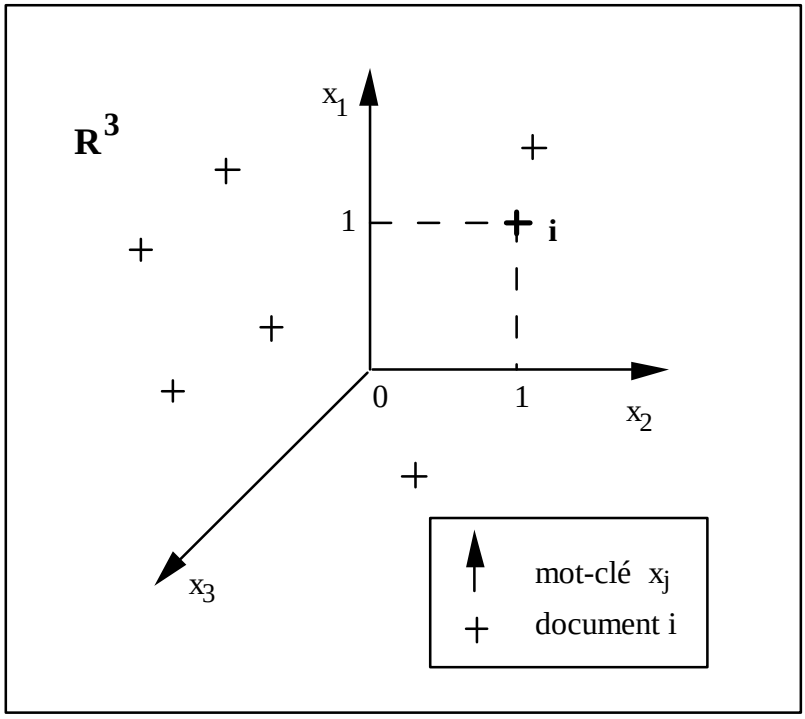


Figure 2 : Représentation d'un corpus documentaire dans un espace géométrique $\mathbf{R^3}$.

Tandis que les techniques de classification non hiérarchiques usuelles représentent les **K** classes recherchées par leur centre de gravité, les **K-means axiales** définissent les **K** classes recherchées par **K** demi-axes passant par l'origine de l'espace géométrique, ou **K** vecteurs unitaires pointant dans la direction de ces demi-axes. La position des K demi axes est initialisée au hasard ou par les K premiers documents. Nous calculons ensuite les **projections orthogonales $y_i(k)$** de chaque document i normé sur les K demi-axes ainsi définis (figure 3), en effectuant les produits scalaires entre le document i normé et les vecteurs unitaires des **K** demi-axes. Chaque document est affecté à la classe k où sa projection y(k) sur l'axe 0Ak est maximale et la position de l'axe est mise à jour[6] pour prendre en compte cette affectation. Par itérations successives, les axes se positionnent puis se stabilisent dans les zones de forte densité du nuage de documents, effectuant ainsi une classification stricte des documents. Pour obtenir des classes recouvrantes, nous définissons ensuite un "seuil de typicité" : un document appartient à la classe si sa valeur de projection sur l'axe représentant la classe est supérieure au seuil. Un document peut donc appartenir à plusieurs classes si ses valeurs de projection sur les axes correspondants sont supérieures au seuil.

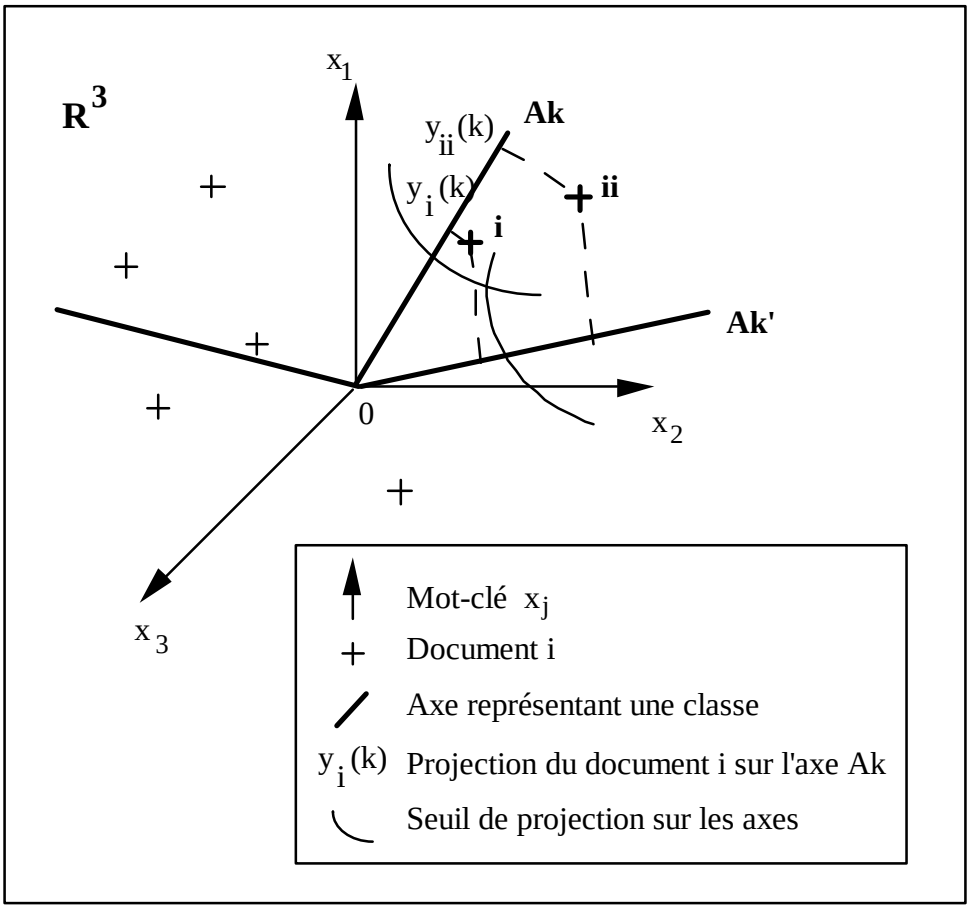


Figure 3 : Partition définitive des documents dans les classes.

[6] immédiatement dans la forme adaptative de l'algorithme, et après passage de tous les documents dans sa forme non adaptative

Sur la figure 3 :

- le document i présente une projection sur l'axe Ak ($y_i(k)$) supérieure au seuil, tandis que sa projection sur l'axe Ak' ($y_i(k')$) est inférieure au seuil. Le document i appartient donc à la seule classe k.
- le document ii présente des projections sur les axes Ak et Ak' supérieures au seuil, le document ii appartient donc aux deux classes k et k'.

Sur la figure 3, nous observons également que la projection du document ii sur l'axe Ak ($y_{ii}(k)$) est supérieure à celle du document i ($y_i(k)$). Nous pouvons donc ordonner les documents appartenant à une classe selon la valeur de leur projection sur l'axe représentant la classe. Cet ordre correspond à un ordre de "typicité" décroissant des documents par rapport au type idéal de la classe qui est un document fictif positionné exactement sur l'axe de la classe dans l'espace géométrique.

En utilisant les valeurs des composantes du vecteur unitaire des classes, nous pouvons définir de la même façon une partition des mots-clés du corpus documentaire. Comme pour les documents, la partition ainsi établie admet des classes recouvrantes, un mot-clé peut appartenir à plusieurs classes, et les mots-clés sont ordonnés selon un ordre de "typicité" décroissant par rapport au type idéal de la classe. La pondération utilisée pour calculer la valeur de "typicité" permet de faire ressortir les mots-clés spécifiques (ou typiques) de la classe, c'est à dire fréquents dans cette classe et rares dans l'ensemble des documents.

Cet algorithme , paramétré par le nombre maximal de classes désiré et le seuil des coordonnées des documents et des mots-clés sur les axes, permet donc de construire des classes d'un type particulier :

- ces classes sont **recouvrantes** car un document ou un mot-clé peut appartenir à plusieurs classes à la fois ;
- les éléments, documents et mots-clés de chaque classe, sont **ordonnés** selon un degré de ressemblance au type idéal de la classe.

Cartographie par **Analyse en Composantes Principales**

Une classe de documents correspond à un **thème**, sous-ensemble homogène de l'information contenue dans le corpus documentaire étudié. Une **Analyse en Composantes Principales** de l'ensemble des classes dans l'espace géométrique permet de déterminer un plan déformant le moins possible le nuage de points de ces classes. Tous les points de ce nuage sont ensuite projetés sur ce plan, constituant ainsi la carte globale des thèmes. Sur cette carte, deux thèmes éloignés représentent des classes dissemblables quant aux mots-clés les définissant. Sur de telles cartes, on peut repérer en particulier des thèmes "exceptionnels", ou des sous-groupes de thèmes.

Complémentarité des méthodes

A. Lelu a démontré que les 2 méthodes sont symétriques l'une de l'autre [page 93, Lelu 93].

- Les *K-Means Axiales* effectuent une classification des documents, en utilisant comme indice de similarité entre documents et classes, le produit scalaire entre les vecteurs documents normés et les vecteurs classes normés [page 72, Lelu 93].
- L'algorithme de classification utilisé par les *Mots Associés* travaille dans l'espace dual de celui présenté dans la méthode des KMeans axiales. Dans cet espace, la cooccurrence entre 2 vecteurs mot-clés I et J correspond au produit scalaire entre I et J. L'indice de similarité utilisé $E_{ij} = C_{ij}^2 / (C_i * C_j) = (C_{ij} / \tilde{A}(C_i * C_j))^2$ correspond, au carré près, à une normalisation de la cooccurrence, c'est à dire au produit scalaire de I et J normés [page 93, Lelu 93].

En résumé, les *K-Means Axiales* effectuent une classification des lignes dans un tableau documents x descripteurs, tandis que les *Mots Associés* effectuent une classification des colonnes de ce même tableau, en utilisant le même indice de similarité[7].

Or, dans nos applications, les tableaux de données sont très creux **et peuvent se segmenter le plus souvent en blocs de lignes et de colonnes quasi-indépendants les uns des autres. Dans ce cas limite, la classification sur les lignes et la classification sur les colonnes aboutissent à détecter les mêmes blocs dans le tableau.**

En effet, dans nos expérimentations, nous n'avons pas relevé de contradictions entre les résultats des deux méthodes sur un même fichier de données. En les paramétrant de façon à obtenir un nombre identique de classes à partir d'un même fichier de données, il est courant d'observer entre 60 et 80 % de classes similaires. Les deux méthodes détectent sensiblement les mêmes blocs. Leur emploi sur un même fichier permet donc d'obtenir des représentations différentes des classes que nous récapitulons ici :

Les classes de mots-clés des *Mots associés* sont structurées par des associations internes et externes. Pour les différencier des demi-axes représentant les classes de mots-clés non structurées et recouvrantes des *Kmeans axiales* , nous les appelons **clusters.** Les clusters sont disjoints, mais on peut considérer que la notion d'association externe adoucit cette classification stricte, de la même manière que la notion de seuil permet d'adoucir la classification des documents par les.*Kmeans axiales.*. Les **clusters** de mots-clés sont relativement faciles à interpréter, car la notion de cooccurrence est intuitivement compréhensible par tout un chacun. Dans les deux cas, les classes de documents sont recouvrantes. Les classes de documents, obtenues par les*Kmeans axiales* sont en général explicites car triées par valeur de projection des documents sur les axes.

Les cartes *des Mots associés*, construites à partir des mesures de centralité et de cohésion des clusters, fournissent une représentation synthétique de la morphologie du réseau. Si ces cartes permettent de comparer les clusters d'un point de vue structurel, elles ne rendent pas compte des proximités entre thèmes comme les cartes par ACP de

[7] L'algorithme de classification utilisé, le simple lien, utilise uniquement l'ordre des paires de mots-clés pour regrouper les mots au sein d'une même classe. Il est invariant par transformation monotone de la matrice de similarités. Aussi du point de vue du résultat de la classification, il est indifférent d'utiliser E_{ij} ou sa racine et donc considérer qu'il s'agit du même indice de similarité.

NEURODOC ou comme pourrait le faire une carte obtenue par “bi-dimensionnal scaling” (coword maps [PETERS et VAN RAAN 93]). C’est pourquoi figurent sur les cartes de SDOC les relations entre thèmes mises en évidence par les associations externes. Les cartes par ACP de NEURODOC, où la distance entre thèmes a un sens d’un point de vue sémantique, sont intuitivement plus lisibles mais nécessitent une certaine expérience pour leur interprétation. En effet, il faut garder à l’esprit que les thèmes les mieux représentés se situent aux extrémités des axes horizontaux et verticaux, ainsi il est possible de dégager des oppositions entre thèmes et par là les grandes lignes d’organisation de ces derniers.

2.2 - Technologie informatique

Nos choix ont visé :

- d’une part, à maîtriser la diversité des méthodes à mettre en oeuvre et des formats bibliographiques existants, ainsi que les volumes d’information à traiter;
- d’autre part, à fournir à l’utilisateur une interface conviviale pour traiter l’information, visualiser et analyser les résultats.

Pour atteindre le premier objectif, nous avons utilisé les techniques du Génie Logiciel : modularité par décomposition en programmes indépendants, adoption de standards. La station de travail a été conçue comme un outil modulaire doté d’un ensemble de fonctionnalités qui peuvent être mises en oeuvre selon les besoins de l’analyse.

Pour atteindre le deuxième objectif, nous avons estimé qu’il fallait avant tout banaliser et standardiser le processus de traitement de l’information en l’automatisant.

a) Une conception modulaire basée sur des standards

La nature textuelle des données à analyser, la diversité de leur structure, le nombre de champs différents à traiter pour mener à bien une étude infométrique, nous ont amenés à adopter la norme SGML[8] pour la description de la structure logique de tous les document manipulés par les outils de la station. Les avantages immédiats de ce choix sont : distinction nette entre contenant et contenu, codage unique des caractères accentués, règles de balisage, existence d’outils sur le marché, ...

A titre d'exemple, une notice bibliographique provenant d'un serveur ou d'un CD-ROM se présente généralement comme suit :

```
NO  : 90-0128293
TI  : Construction automatique de liens hypertextes
AU  : FLUHR (C.)
FD :  Representation connaissances;Lien, Hypertexte;
...
...
```

[8] SGML : Standard Generalized Mark-up Language.

La structure logique d'une telle information est très simple : une suite de champs repérés par un identifieur. Il est alors facile de définir les règles lexicales qui permettent d'identifier le début, la fin d'une notice, le début ou la fin d'un champ à l'intérieur de la notice de manière à la transformer en document SGML.

En SGML, chaque élément structurel est repéré par une balise de début : <identifieur de l'élément> et une balise de fin : </identifieur de l'élément>. La notice ci-dessus peut d'écrire en format SGML :

```
<record>
<NO>90-0128293</NO>
<TI>Construction automatique de liens
hypertextes</TI>
<AU>FLUHR (C.)</AU>
<FD>Representation connaissances;Lien; Hypertexte;
...</FD>
...
</record>
```

Une fois que toutes les données sont décrites dans ce format pivot, il est plus facile de concevoir des outils génériques utilisant les propriétés du balisage SGML. La plupart des traitements sur de tels documents se réduisent à associer des actions à un élément de la grammaire et, dans bien des cas, travailler au niveau lexicographique suffit. Ces caractéristiques nous ont conduits à développer une boite à outils (appelée ILIB) basée sur SGML et sur UNIX [DUCLOY *et al* 1991]. En effet des programmes générés par Lex et des outils UNIX tels que Awk sont bien adaptés pour extraire de l'information "à la volée" sur un flot de données structurées, puis la traiter.

La station de travail est ainsi constituée de modules indépendants de traitement de l'information qui communiquent entre eux par flot de données en s'appuyant sur le mécanisme de *pipe* d'UNIX. En collaboration avec H. Millerand et J. Kasprzak du service étude de la direction informatique INIST, nous avons effectué des tests d'applications de SDOC et NEURODOC sur de gros volumes de données. (transcrits dans le guide technique de SDOC et NEURODOC). A titre d'exemple, le traitement de 16 000 références bibliographiques par l'un ou l'autre des outils prend environ dix heures sur une machine déjà ancienne, Sun Sparc 1, avec 16 Mo de mémoire vive. Il faut noter que ce n'est pas la phase de classification elle-même qui est longue, mais la phase de documentation des classes (libellés des mots-clés, titres, sources, auteurs, ...); celle-ci prend plus de la moitié du temps d'exécution. Elle sera optimisée ultérieurement.

b) Interface utilisateur : Scénarii d'analyse standard et mise en forme des résultats

Dans le souci de faciliter l'utilisation de cette station de travail, nous avons défini des scénarii d'analyse standards. Ces derniers sont matérialisés par des "fichiers de paramètres standards" où sont définis les paramètres de l'analyse (directement dépendants de la méthode choisie) et les différentes éditions ou mises en forme de résultats souhaitées. L'utilisateur peut donc éditer un fichier de paramètres standard, le modifier, l'enregistrer sous un autre nom, puis demander l'exécution de telle ou telle phase de traitement à partir du nouveau fichier de paramètres.

Nous avons apporté un soin particulier à la mise en forme des résultats avec comme objectif d'obtenir des représentations lisibles et combinables favorisant l'intuition et les rapprochements d'idées. Pour cela, nous nous sommes appuyés sur trois techniques :

- le transfert des résultats vers des applicatifs spécialisés (tableurs, éditeurs, ...). Exemples tableaux 1 et 2,
- les langages de composition (code interprété par un logiciel ou une imprimante) tels que PostScript , nroff, troff et LaTex[9] pour les éditions de documents que nous avons désiré automatiser complètement,
- l'hypertexte[10] pour la navigation dans l'espace documentaire constitué des cartes thématiques, classes de mots-clés et de documents, liste d'auteurs, ...). [GRIVEL et LAMIREL 1993], [LELU et FRANCOIS 1992]. Exemples : figures 5, 6, 7, 8.

2.3 - La chaîne de traitement infométrique [POLANCO et al. 1993a]

La figure 4 présente le déroulement général d'une application scientométrique.

---

[9] PostScript® est une marque déposée de Adobe. nroff et troff sont des formatteurs de texte disponibles en standard sous UNIX. LaTeX est un environnement (langage et programme) bâtit sur TeX, marque déposée de American Mathematical Society, disponible par ftp : ftp.inria.fr/TeX/

[10] Un document hypertexte est un fichier de texte où figurent des liens vers d'autres parties du document lui-même ou vers d'autres documents. La présence de liens dans un document est mise en évidence par une signalétique pré-définie (boutons, mots en gras ou encadrés, ...). Cela signifie, qu'en cliquant sur ces zones (appelées également ancres), on accède à un autre document. Dans notre cas, les documents ne contiennent pas seulement du texte mais aussi des images (cartes thématiques). Ce sont des documents hypermedia.

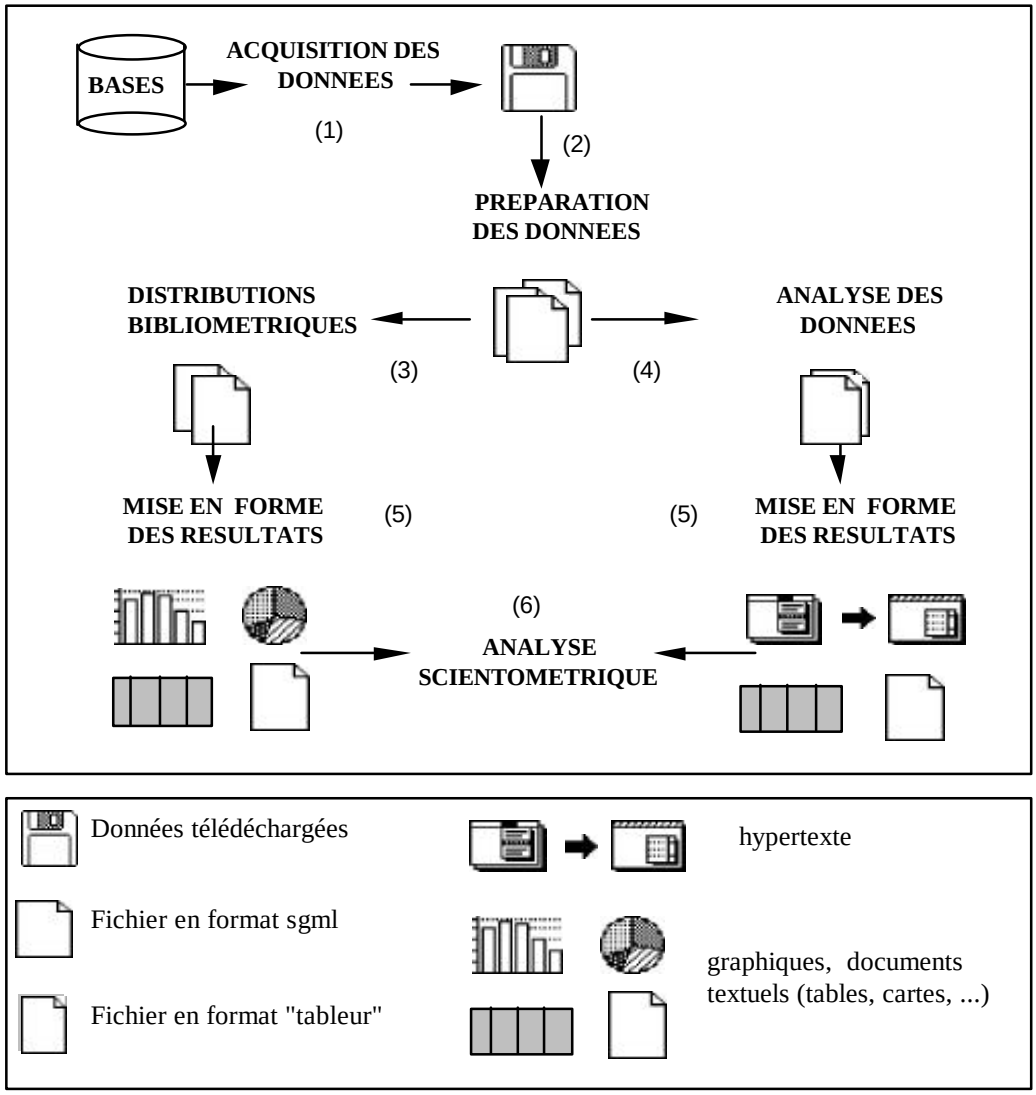


Figure 4 : la chaîne de traitement infométrique.

Le schéma de traitement proposé comprend 6 phases successives : 1) acquisition des données à analyser, 2) préparation des données, 3) distributions bibliométriques, 4) analyse des données, 5) mise en forme des résultats, 6) analyse scientométrique des résultats.

Les phases 2 à 5 sont automatisées et seront décrites dans ce paragraphe. La phase d'acquisition des données (1) est manuelle et dépend des données à étudier; elle ne sera pas détaillée ici. L'analyse scientométrique des résultats (phase 6) est manuelle; elle est traitée dans la deuxième partie de l'article.

La préparation des données (phase 2) :

Cette phase permet de normaliser la collection de documents (reformatage) et de générer les données nécessaires aux phases 3 et 4, à savoir :

- la collection de documents initiaux convertis au format SGML;

- des index qui permettent de repérer pour chaque forme[11], tous les endroits du corpus où elle est utilisée. Par exemple, un index des mots-clés associe à chaque mot-clé la liste des numéros des documents qu'il indexe.

Cette phase est paramétrée par le noms des champs bibliographiques pour lesquels la création d'index est effectué ainsi que par les séparateurs de forme. Les index sont également des documents SGML.

Des distributions bibliométriques (phase 3) :

Cette phase a pour objectif d'établir les distributions des champs à étudier. Outre des informations de nature quantitative sur les données, elle fournit des indicateurs utilisés pour le paramétrage de la classification. Elle est également paramétrée par le nom des champs sur lesquels les comptages sont effectués.

L'analyse des données (phase 4) :

Cette phase permet la création des classes de mots-clés et de documents en format SGML. Elle est indépendante du format initial des documents, car elle prend en entrée les données normalisées obtenues par la phase 2.

Deux logiciels sont disponibles à ce jour :
- le logiciel SDOC (implémentation de la méthode des mots associés),
- le logiciel NEURODOC (implémentation des K-means axiales et d'une Analyse en Composantes Principales).

Les traitements de SDOC s'effectuent en 4 étapes : 1) calcul des cooccurrences de mots-clés et mesure de la force d'association des paires de mots-clés, 2) classification : regroupement des mots-clés en clusters, 3) calcul des coordonnées géographiques des clusters, 4) affectation aux clusters des documents et des informations relatives à ceux-ci (titre, auteurs, sources).

Les traitements de NEURODOC s'effectuent en 3 étapes : 1) calcul des classes de mots-clés et de documents par la méthode des K-means axiales, 2) calcul des coordonnées géographiques des classes sur un plan par une Analyse en Composantes Principales, 3) documentation des classes, c'est à dire addition du libellé des mots-clés, du titre des documents, des auteurs et des sources associés.

La mise en forme des résultats (phase 5) :

Cette phase permet à l'utilisateur de visualiser les résultats des phases 3 et 4. Les représentations générées sont les instruments de travail de l'analyse scientométrique (phase 6).

3 - Analyse scientométrique des résultats

[11] suite de caractères encadrée par un caractère jouant un rôle de séparateur [LEBART et SALEM 1988]

3.1 - Exploitation des distributions bibliométriques

A partir des différentes distributions, plusieurs types d'observations peuvent être effectuées. Pour un domaine donné, on peut ainsi quantifier sa magnitude (nombre d'articles, nombre de revues), son actualité (selon la date de publication), sa localisation (selon le pays d'édition des revues scientifiques), l'importance des périodiques scientifiques (selon le nombre d'articles dont ils sont la source au cours d'une période déterminée), la localisation des auteurs (selon leur appartenance institutionnelle) et son vocabulaire d'indexation.

Tous ces éléments seront également utilisés pour orienter une analyse approfondie d'un domaine particulier. Ils permettront de définir un corpus de références bibliographiques homogène et pertinent, sur lequel les méthodes d'analyse des données peuvent être appliquées. Par exemple, on peut utiliser la loi de Bradford pour focaliser son attention sur les revues les plus "productives" en termes d'articles recueillis dans le corpus, ainsi que la loi de Zipf pour déterminer le vocabulaire d'indexation pertinent pour l'analyse. Cette loi nous permet de séparer le vocabulaire d'indexation en trois groupes :
- un ensemble restreint de mots-clés de fréquence élevée mais trop généraux (information triviale);
- un ensemble de mots-clés de fréquence plus faible mais riches en information;
- un ensemble très important de mots-clés de fréquence très faible (1 ou 2), difficile à exploiter d'un point de vue statistique et générateur de bruit (information marginale)

C'est donc le second ensemble de mots-clés qui fournit l'information la plus intéressante et qui est traité par les méthodes d'analyse de données.

3.2 - Exploitation des résultats des méthodes d'analyse de données

Pour chaque méthode, nous décrirons la structure des classes obtenues, puis le protocole d'interprétation des classes et cartes. Celui-ci est basé sur l'observation d'indicateurs générés automatiquement permettant d'apprécier la qualité de la classification obtenue d'un point de vue global puis local à chaque classe. Nous suivrons un plan rigoureusement parallèle pour permettre une comparaison entre les deux méthodes. Nous utiliserons les résultats du traitement d'un corpus de références extraites de la base PASCAL, au début de l'année 1990, dans le domaine des Sciences de l'ingénieur : "Intelligence Artificielle : systèmes experts". Ce corpus comprend 316 références, il est indexé par 955 mots-clés dont 665 de fréquence 1 (soit 70% du vocabulaire d'indexation).
La première étape des deux analyses présentées ci-dessous a consisté en une sélection du vocabulaire d'indexation en se basant sur la loi de Zipf:
- suppression des 4 mots-clés le plus fréquents : Intelligence artificielle, Système expert, Base de connaissance et Représentation des connaissances;
- suppression des mots-clés de fréquence 1.

3.2.1 - Analyse des résultats fournis par SDOC

Le résultat de la classification est une **partition** des mots-clés en classes structurées mais disjointes (clusters), même si les clusters peuvent entretenir des relations avec d'autres clusters. Un cluster représente un thème trouvé dans un ensemble de documents.

a) Anatomie des clusters

La figure 5 décrit l'un des 21 clusters obtenus par SDOC sur ce corpus en limitant la taille des clusters à 10 mots et en fixant une cooccurrence minimale des mots-clés à 2.

Baia navigation : cl14 (shuu.FR)

Noeud Options Editeur Précédent

SDOC

home | 15 titres | 25 auteurs | 6 sources | carte

Thème : Recherche documentaire

- Donneés quantitatives

Coefficient de saturation = 0.10
Nombre de citations par les autres clusters = 4
Indice de cohésion = 0.24
Indice de centralité = 0.06

- Mots-clés du thème

Poids	Fréquence	Mots-clé
0.60	8	Recherche documentaire
0.30	8	Traitement automatisé
0.20	5	Analyse sémantique
0.10	3	Analyse linguistique
0.10	3	Texte intégral
0.10	2	Formulation question
0.10	4	Indexation
0.05	2	Concept
0.05	3	Analyse syntaxique
0.10	16	Conception
0.05	32	Modélisation
0.05	18	Conception assistée
0.05	6	France
0.05	6	Interface
0.05	6	Traitement en ligne
0.05	4	Système documentaire

Baia navigation : cl14 (shuu.FR)

Noeud Options Editeur Précédent

- Associations internes

Poids	Coocurrence	Association
0.67	2	Formulation question<->Texte intégral
0.38	3	Analyse linguistique<->Traitement automatisé
0.27	2	Analyse sémantique<->Analyse syntaxique
0.25	4	Recherche documentaire<->Traitement automatisé
0.25	2	Formulation question<->Recherche documentaire
0.25	2	Concept<->Recherche documentaire
0.20	2	Analyse sémantique<->Indexation
0.17	2	Recherche documentaire<->Texte intégral
0.17	2	Analyse linguistique<->Recherche documentaire
0.12	2	Indexation<->Recherche documentaire
0.10	2	Analyse sémantique<->Traitement automatisé
0.10	2	Analyse sémantique<->Recherche documentaire

- Associations externes

* Avec le thème Hypertexte :

0.12	2	Recherche documentaire<->Système documentaire
0.03	2	Traitement automatisé<->Conception
0.03	2	Recherche documentaire<->Conception

* Avec le thème Interface utilisateur :

0.08	2	Recherche documentaire<->Traitement en ligne

* Avec le thème Langage naturel :

0.08	2	Recherche documentaire<->Interface
0.08	2	Recherche documentaire<->France

* Avec le thème Conception assistée :

0.03	2	Traitement automatisé<->Conception assistée

Figure 5 : Exemple de thème obtenu avec SDOC : "Recherche documentaire"

Un cluster est composé de :

- une liste de mots-clés,
- une liste d'associations internes,
- une liste d'associations externes,
- une étiquette,
- une liste de documents affectés après la classification.

La liste de mots-clés regroupe des mots qui sont proches les uns des autres. Nous distinguons les mots clés internes (qui apparaissent dans les associations internes) des mots-clés externes (qui apparaissent seulement dans les associations externes car ils ont été rejetés de ce cluster à cause du critère de taille maximal des clusters). Ainsi, sur la

figure, les mots-clés figurant dans les associations internes constituent les mots clés internes du cluster Recherche documentaire et les mots-clés situés à droite dans les associations externes constituent les mots-clés externes du cluster. Par exemple Interface dans Recherche documentaire - Interface sera l'un de ses mots-clés externes. Les mots-clés sont triés selon leur nombre d'apparitions dans les associations internes et externes du cluster.

La liste d'associations internes décrit la force des associations des mots qui définissent la structure interne des clusters. Par exemple, l'association Analyse sémantique - Analyse syntaxique du cluster Recherche documentaire a un poids de 0.27. Plus la valeur de l'association est forte, plus les mots sont fortement associés.

La liste d'associations externes décrit les associations existants entre les mots d'un cluster et les mots d'autres clusters. Dans l'exemple de la figure 5, l'association Recherche documentaire - Interface relie les clusters Recherche documentaire et Langage naturel. Le nombre d'associations externes peut être limité aux N plus fortes. Dans ce cas, les associations externes ne sont pas nécessairement bi-directionnelles. Dans le cas présent, nous l'avons limité aux 10 plus fortes.

Etiquetage des clusters : le choix d'un terme représentatif pour nommer le cluster est basé sur une heuristique. Nous choisissons le terme de la liste des mots-clés internes qui apparaît le plus grand nombre de fois dans les associations internes et externes. Par exemple, le programme SDOC proposera le mot-clé Recherche documentaire pour désigner le cluster de la figure 5. Le nom proposé est satisfaisant dans plus de 90% des cas.

La liste des documents affectés à un cluster : elle est obtenue après exécution de la classification. C'est la liste des documents qui ont contribué à la formation de ce cluster par la présence dans leur indexation de couples de mots-clés qui pourraient constituer une association interne ou externe du cluster. Un document peut donc figurer dans plusieurs clusters. Un document ne figurant que dans un seul cluster est appelé document propre au cluster. Les documents sont triés selon l'importance de leur contribution à l'élaboration du cluster. A partir des documents sont extraits le titre, les auteurs et la source pour compléter la description du cluster.

b) Interprétation de la partition obtenue

- Qualité de la partition

Des indicateurs globaux permettent d'apprécier la validité du paramétrage et caractérisent la partition.

- le nombre de documents et de mots-clés classés permet de mesurer **la "réduction" des données**, c'est à dire la part d'information contenue dans le corpus étudié mais perdue dans la partition obtenue ; dans notre exemple nous avons conservé 199 documents dans les classes (environ 2/3 de l'information bibliographique initiale), et 149 mots-clés (15% du vocabulaire d'indexation initial); Ce résultat plutôt faible concernant le pourcentage d'information bibliographique présent de la partition s'explique par le fait que le seuil de cooccurrence choisi (2) élimine 88 documents

sur 316 et 757 mots-clés sur 955. La classification elle-même a peu d'influence concernant la perte d'informations. En général, on cherche à obtenir 80 % des documents avec environ 20% des mots-clés.

- Le nombre d'occurrence de documents dans les clusters (dans notre cas 321) doit être examiné à la lumière de la distribution des documents dans les clusters. Celle-ci a un comportement analogue à la loi de Zipf. 53% des documents classés ne sont présents que dans un seul cluster, 30% dans deux clusters, 10 % dans 3 clusters, etc. Ces chiffres permettent d'évaluer le niveau d'inclusion mutuelle ou recouvrement des ensembles de documents associés aux clusters. Ce **taux de recouvrement** des classes de documents est en partie maitrisable par l'utilisateur en limitant le nombre d'associations externes aux N plus fortes.

- Caractéristiques des clusters

Un tableau résumant les caractéristiques structurelles des clusters permet de les catégoriser et d'apprécier la répartition des documents dans les clusters.

Nom	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Revetement metallique	0.66	0.75	0.00	5	0	8	0	0	3	3
Industrie bâtiment	0.40	0.62	0.00	8	0	20	0	11	4	1
...										
Recherche documentaire	0.10	0.24	0.06	9	7	12	8	4	15	5
Langage naturel	0.06	0.20	0.04	6	5	6	8	3	14	4
Processus acquisition	0.04	0.07	0.04	8	10	9	10	6	27	8
...										

Tableau 1 : Exemple de tableau des caractéristiques des clusters

[1]:Seuil de saturation, [2]:densité, [3]:centralité, [4]:Nombre de mots-clés internes, [5]:Nombre de mots-clés externes, [6]:Nombre d'associations internes, [7]:Nombre d'associations externes avec d'autres clusters, [8]:Nombre de citations du cluster par d'autres clusters, [9]:Nombre de documents définissant le cluster, [10]:Nombre de documents propres au cluster.

Le seuil de saturation d'un cluster [1] est la valeur de la dernière association interne ajoutée avant sa saturation, c'est à dire lorsqu'il ne peut plus grandir en taille. Trier le tableau selon cette valeur permet de connaître l'ordre dans lequel les clusters se sont figés. Ainsi, le cluster Langage naturel s'est stabilisé après le cluster Recherche documentaire.

La densité [2] d'un cluster est la moyenne des associations internes du cluster. C'est un indicateur de sa cohésion, son homogénéité. L'examen de sa taille[4] et de son nombre d'associations internes [6] permet d'avoir une idée plus précise de cette cohésion. La densité de Recherche documentaire est presque similaire à celle de Langage naturel mais le rapport "nombre de mots qui le constituent" sur "le nombre de connections entre ces mots" est plus faible, indiquant une connectivité plus importante. On peut dire que Recherche documentaire a une cohésion plus forte que Langage naturel. La somme des valeurs de [4] donne le nombre de mots-clés gardés dans les clusters.

La centralité d'un cluster [3] est la valeur moyenne des associations externes. Le nombre de citation [8] d'un cluster indique le nombre de fois qu'un cluster est cité par les autres clusters via leurs associations externes. On considère que les colonnes [3], [5], [7] and [8] caractérisent les associations externes d'un cluster et permettent d'apprécier son rattachement au réseau. Ainsi les 2 clusters Recherche documentaire et Langage naturel ont de nombreux liens avec les autres clusters du réseau, tandis que Revetement metallique est particulièrement isolé. Le cas de Industrie bâtiment est un petit peu plus complexe car il n'a pas d'associations externes mais est cité 11 fois. La navigation hypertexte permet de lever immédiatement ce mystère en facilitant l'accès à la description des clusters. En fait, il existe un thème nommé Conception assistée traitant des applications de l'IA dans l'industrie naval qui fait neuf fois référence à Industrie bâtiment à travers le terme Conception assistée. On a donc en réalité deux thèmes autonomes : Industrie bâtiment et un thème qu'on peut appeler Industrie naval, aux vocabulaires très spécifiques reliés par un terme plus générique de fréquence plus élevée Conception assistée. Le tri du tableau complet des clusters par centralité permet de situer la force de ces liens qui dans le cas présent était relativement élevée pour Recherche documentaire (dans le premier tiers d'un tableau de 21 clusters).

Les colonnes [9] et [10] permettent d'apprécier la répartition des documents dans les clusters. Comme un document peut appartenir à plusieurs clusters, le nombre total de document classés dans un cluster donné [9] est distinct du nombre de documents propres au cluster [10]. Aussi la somme des valeurs de la colonne [9] donne le nombre d'occurrences de documents dans les clusters. La somme des valeurs de la colonne [10] donne le nombre de documents qui ne figurent que dans un seul cluster. Le rapport des colonnes [9] et [10] donne le pourcentage de documents propres à un cluster.

Nous utilisons une catégorisation des clusters décrite dans [COURTIAL 1990, page100] pour définir un plan de lecture des clusters. Un cluster est dit **principal** si son seuil de saturation [1] est plus élevé que celui de ces clusters associés ou clusters externes. L'intensité de ses associations externes [3] est généralement inférieure à son seuil de saturation. Les clusters associés sont appelés clusters **secondaires**. Ils sont l'extension naturelle du cluster principal. Ainsi Recherche documentaire est un exemple de cluster principal avec comme cluster secondaire associé Langage naturel qui par ailleurs joue un rôle de cluster principal vis à vis de processus acquisition. Par cette méthode de lecture, le découpage en classes de taille fixes ne change pas les résultats que l'on cherche à mettre en évidence.
Dans une lecture des clusters en vue d'une analyse, nous privilégions les clusters principaux entretenant de nombreuses relations avec d'autres clusters, en vue d'appréhender le plus rapidement possible les principaux noeuds thématiques du réseau.

Pour établir ce plan de lecture, le tableau des caractéristiques des clusters ne suffit pas. Il faut également utiliser la description complète des clusters, en particulier étudier précisément leurs associations externes pour les situer les uns par rapport aux autres, comme on l'a vu par exemple dans le cas du cluster Industrie Batiment.

c) Cartographie

Les cartes fournissent une synthèse visuelle de deux paramètres du tableau précédent : la densité et la centralité. Nous utilisons ce mode de représentation pour obtenir une carte par cluster, avec visualisation de ses relations s'il possède des associations externes. Pour éviter le recouvrement des clusters ayant des coordonnées voisines, on peut redéfinir ces coordonnées en les classant selon leur rang. C'est la technique employée pour la figure 6.

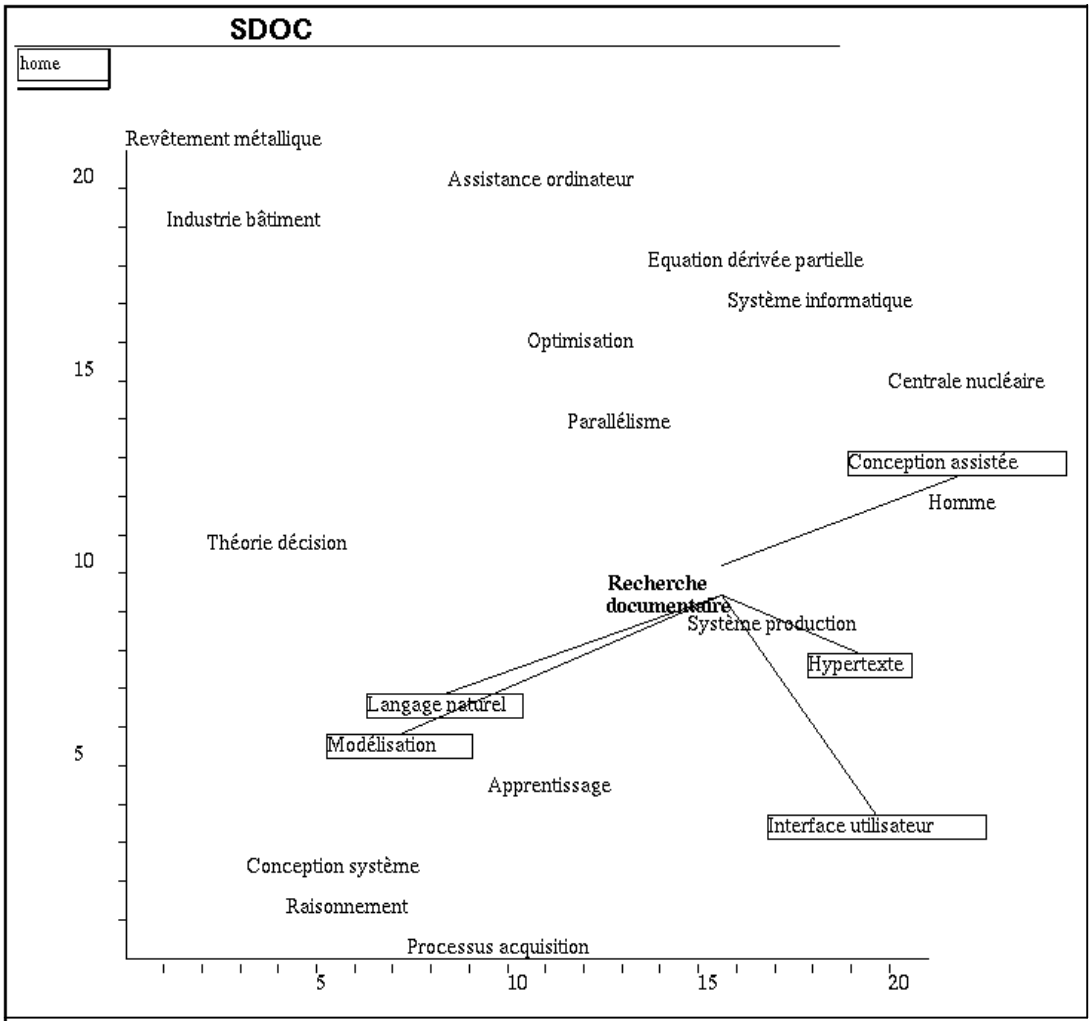


Figure 6 : Exemple de carte thématique obtenue avec SDOC

- interprétation des cartes

Nous resterons dans notre lecture de la carte au niveau d'une explication des résultats à partir du corpus étudié et de son vocabulaire d'indexation, sans faire d'interprétation sur l'intéret stratégique des thèmes, type d'interprétation qui n'est pas de notre compétence. Puis nous montrerons que la visualisation des noms des clusters et la mise en évidence graphique des relations existants entre clusters peut permettre à un utilisateur de focaliser son attention sur un thème particulier et d'examiner des sous-réseaux du réseau global.

Dans l'exemple de la figure 6, on peut relever que les cluster Revêtement métallique et Industrie bâtiment sont a priori isolés par rapport au corpus (forte densité, faible centralité), ce qui est confirmé par le nombre et l'examen des documents associés. Les

thèmes à forte densité se situent dans la partie haute de la carte. Ce sont ici des applications de l'intelligence artificielle (revêtement métallique, industrie bâtiment, systèmes experts pour la résolution d'équation à dérivées partielles, domaine documentaire, ...). Les documents en question ont une indexation très spécifique pour décrire le domaine d'application. Les thèmes situés au bas de la carte ont une cohésion plus lâche. Ils correspondent ici en général à des thèmes plus théoriques de l'intelligence artificielle (raisonnement, modélisation, apprentissage, etc). Ils sont constitués de mots à fréquence élevée et regroupent des ensembles de documents plus importants que les précédents.

Si on se focalise sur un thème particulier, comme ici Recherche documentaire , on peut examiner son réseau local. Etant donné le corpus étudié, il n'est pas surprenant de trouver de grands types d'application de l'IA à l'informatique documentaire tels que les interfaces évoluées (hypertexte), les systèmes d'analyse linguistique (langage naturel), les systèmes experts fondés sur une représentation conceptuelle de documents (un sous-thème présent dans le cluster modélisation). La liaison avec Conception assistée exprime elle une relation plus générale entre les mots-clés "traitement automatisé" et "Conception assistée" sans qu'il y ait de rapports directs avec la recherche documentaire. En effet le cluster "Conception assistée" traite en fait d'applications de l'IA dans la construction navale. La navigation hypertexte permet de suivre les associations intéressantes et les cartes sont d'un grand secours pour éviter de se perdre au cours de la consultation.

3.2.2 - Analyse des résultats fournis par NEURODOC

Le résultat de la classification est une **partition** des mots-clés et des documents en classes recouvrantes. Une classe ainsi définie correspond à un thème, sous-ensemble homogène de l'information contenue dans le corpus documentaire étudié.

a) Anatomie des classes obtenues

La figure 7 montre l'exemple de la classe ou du thème "Hypertexte" tel qu'il apparaît dans un des dispositifs hypertextes possibles (le logiciel Hypercard® sur Macintosh®[12]). Un thème est donc constitué de quatre listes : mots-clés, documents, auteurs et sources triées par ordre de pertinence décroissant par rapport au type idéal de la classe.
Une classe est nommée par le mot-clé de "typicité"[13] la plus forte par rapport au type idéal de la classe (cf § 2). Dans environ 20% des cas, la révision de ce nom par un expert peut être nécessaire.

Un mot-clé est représenté par son libellé et sa valeur de "typicité" par rapport au thème. Les valeurs de"typicité" des mots-clés permettent de distinguer les mots-clés importants pour l'interprétation du thème, et d'estimer la structure de la classe. En effet, nous observons deux types de classes :
- classe dont la typicité des mots-clés décroit de façon continue dans la liste des mots-clés;

---

[12] Macintosh® et Hypercard® sont des marques déposées de Apple Computer Inc.
[13] Nous rappelons que la pondération utilisée pour calculer la valeur de "typicité" permet de faire ressortir les mots-clés fréquents dans cette classe et rares dans l'ensemble des documents.

- classe où nous observons des ruptures importantes dans les valeurs de “typicité”; dans ce cas un nombre restreint de mots-clés définissent le thème. Les classes construites à partir d’un petit nombre de documents présentent donc des mots-clés de “typicité” élevée.

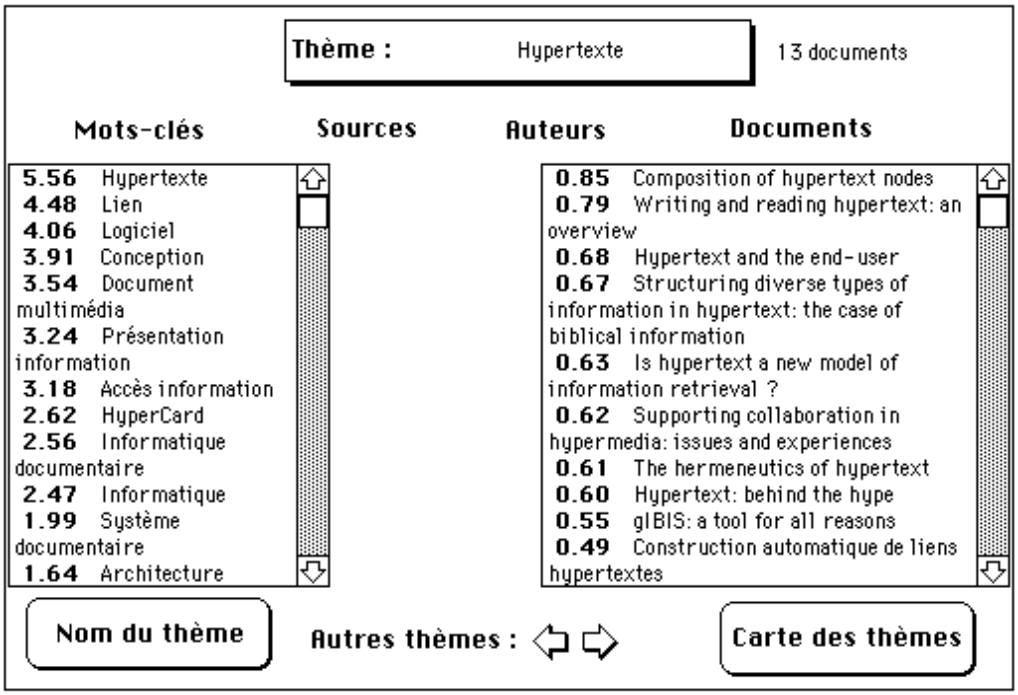


Figure 7 : Exemple de thème obtenu avec NEURODOC

Un document est représenté par son titre et sa valeur de “typicité” par rapport au thème les documents les plus pertinents du thème sont en général les plus spécifiques au thème. Les documents de pertinence moindre se retrouvent dans d’autres thèmes, où ils sont d'ailleurs souvent mieux situés. Dans le dispositif hypertexte, chaque titre de document donne accès à la référence complète.

A partir des documents associés au thème, sont extraits, s’ils existent, les auteurs et les sources de ces derniers. Les auteurs et les sources sont affectés du poids du document correspondant. Si un auteur ou une source est associé à plusieurs documents du thème, les poids de ces derniers sont sommés. Les thèmes sont complétés par la liste triée des auteurs et des sources. Les listes des auteurs et des sources sont visualisables en sélectionnant les mots “auteurs” et “sources”, elles permettent de connaître les équipes de scientifiques les plus importantes pour un thème donné et les principales revues qui publient ces articles.

b) Interprétation de la partition obtenue

- Qualité de la partition :

La classification est effectuée par approximations successives, aussi le récapitulatif du déroulement de la classification permet de vérifier la convergence du processus. Si la

stabilisation n'a pas lieu, il peut être intéressant d'augmenter le nombre de classes pour créer des classes spécifiques aux documents oscillants entre deux classes.

Les indicateurs globaux permettant d'apprécier la qualité de la partition obtenue sont :

- le nombre de classes obtenues : le nombre de classes demandées est un nombre maximal, certains axes initialisés peuvent ne pas avoir été utilisés pour la classification ;

- le nombre de documents et de mots-clés classés permet de mesurer **la "réduction" des données**, c'est à dire la part d'information contenue dans le corpus étudié mais perdue dans la partition obtenue ; dans notre exemple nous avons conservé 250 documents dans les classes (80% des documents traités), et 248 mots-clés (26% des mots-clés totaux); Ce chiffre faible s'explique par le fait que la classification n'est effectuée qu'avec les mots-clés de fréquence > 1, soit 30 % des mots-clés totaux.

- le nombre d'occurrences de documents ou mots-clés obtenus dans l'ensemble des classes, complété par les distributions des documents ou mots-clés dans les classes permet de mesurer le **taux de recouvrement** des classes. Dans notre exemple, nous obtenons 321 occurrences de documents. Sur 250 documents classés, 70% sont spécifiques d'une classe, les 30% restants figurant dans leur quasi totalité dans deux classes. De même sur 248 mots-clés, environ 60% sont spécifiques d'une classe, les 40% restants figurant dans leur quasi totalité dans deux classes.

Ces indicateurs montrent que la réduction des données est du même ordre de grandeur que celle obtenue avec SDOC. Pour l'outil NEURODOC, le taux de recouvrement est dépendant des paramètres de la classification (nombre de classes demandés et seuil des documents et mots-clés). Il est donc maitrisable par l'utilisateur. Dans cet exemple, le taux de recouvrement est suffisamment faible pour considérer les documents et mots-clés conservés dans les classes comme pertinents.

- Caractéristiques des classes :

Un tableau résumant les caractéristiques des classes permet d'apprécier la qualité de la répartition des documents dans les classes et de catégoriser ces dernières.

Dans ce tableau, chaque classe est caractérisée par :

- [1] une valeur d'inertie indiquant la dispersion des documents autour de l'axe représentant la classe, elle correspond à la somme des carrés des projection des documents ayant constitué la classe sur l'axe représentant cette dernière.;
- [2] le nombre de documents ayant construit la classe, c'est à dire le nombre de documents ayant leur projection maximale sur l'axe représentant cette classe ;
- [3] le nombre de documents affectés à cette classe, c'est à dire le nombre de documents dont la coordonnée sur cet axe est supérieure au seuil défini par l'utilisateur ;
- [4] le nombre de mots-clés affectés à cette classe, c'est à dire le nombre de mots-clés dont la coordonnée sur cet axe est supérieure au seuil défini par l'utilisateur ;
- [5] le nombre d'auteurs associés à cette classe ;

- [6] le nombre de sources associés à cette classe.

Nom	[1]	[2]	[3]	[4]	[5]	[6]
Apprentissage	11.24	33	33	12	66	16
Raisonnement	8.88	30	29	17	48	10
Processus acquisition	7.30	19	29	10	56	14
Informatique biomedicale	5.95	19	18	18	54	14
Base donnee	5.63	20	18	21	34	9
Conception assistee	5.08	21	18	23	31	11
Interface utilisateur	4.98	22	21	24	43	7
Assistance ordinateur	4.84	19	19	22	46	14
...						

Tableau 2: Exemple de tableau des caractéristiques des classes

Dans ce tableau, les classes sont triées par valeur d'inertie décroissante. Les premiers thèmes sont généralement les plus importants en taille (colonnes [2] et [3]), ils regroupent les thèmes essentiels du corpus étudié. Pour un nombre de documents égal, plus l'inertie d'une classe est importante, plus les documents constituants sont regroupés de façon pertinente. Par exemple, le thème "Processus acquisition" ([1] = 7,30 ; [2] = 19) regroupe des documents plus homogènes que le thème "Informatique biomedicale" ([1] = 5,95 ; [2] = 19).

Pour apprécier la qualité de la répartition des documents dans les classes, un premier critère est le nombre de documents ayant construit la classe [2]. Si quelques classes regroupent l'essentiel des documents, et si elles correspondent à des mots-clés de très forte fréquence, elles risquent de masquer une information plus pertinente. Aussi, il peut être intéressant d'éliminer ces mots-clés de l'indexation. Dans l'exemple du tableau 2, les deux premières classes regroupent chacune 33 et 30 documents, ce qui est à peine supérieur aux classes suivantes; nous pouvons considérer que les documents sont équitablement répartis.

Une comparaison entre le nombre de documents ayant construit la classe [2] et le nombre de documents affectés à cette classe [3] permet d'estimer la pertinence du seuil des documents:

- si [2] < [3] : la classe regroupe des documents ayant construits d'autres classes et également bien représentés dans cette classe (exemple : thème "Processus acquisition") ;
- si [2] > [3] : certains documents ayant construit cette classe ont une valeur de projection inférieure au seuil, il sont donc perdus lors de la classification (exemple : thème "Raisonnement").

La colonne [4] permet d'estimer la pertinence du seuil des mots-clés. Nous remarquons qu'un thème homogène (exemple : "Processus acquisition", [4] = 10) est défini par moins de mots-clés qu'un thème plus dispersé (exemple : "Interface utilisateur", [4] = 24).

Les colonnes [5] et [6] permettent d'estimer la dispersion des auteurs et des sources (titres des revues) autour des thèmes.

c) Cartographie

Afin de positionner les thèmes obtenus les uns par rapport aux autres, nous représentons les classes obtenues par des points. Une Analyse en Composantes Principales de l'ensemble des points représentant les classes permet de déterminer un plan déformant le moins possible le nuage de points ainsi défini. Tous les points de ce nuage sont ensuite projetés sur ce plan, constituant ainsi la carte des thèmes.Dans le cas présent, nous avons utilisé les coordonnées réelles des thèmes et non le classement par rang, considérant que la carte obtenue (figure 8) restait lisible.

- Interprétation de la carte obtenue (figure 8) :

Sur la carte, la proximité entre deux thèmes indique qu'ils sont définis par des mots-clés issus de domaines connexes. Par exemple, les thèmes : "Hypertexte" et "Interface Utilisateur" sont proches sur la carte, les travaux sur les hypertextes correspondent à un sous-ensemble des problèmes d'interface utilisateur.

La position des thèmes sur la carte est interprétée en fonction des axes horizontaux et verticaux définissant le plan. Dans un premier temps, il est important de garder à l'esprit que les thèmes les mieux représentés sur cette carte se situent plutôt vers les extrémités des deux axes, c'est à dire vers les bords gauche et droit puis haut et bas de la carte. La position des thèmes situés vers le centre de la carte est moins significative.

La carte (figure 8) montre que sur l'axe horizontal s'opposent :

- vers la gauche les thèmes théorique de l'Intelligence Artificielle comme "processus acquisition", "raisonnement", et " méthodologie" ;
- vers la droite, les thèmes applicatifs dans les domaines documentaire ("recherche documentaire"), et interface utilisateur ("base donnée", "interface utilisateur" et "hypertexte").

Sur l'axe vertical:

- s'isolent en haut à gauche les thèmes "automatisation" et "système production" qui correspondent à des applications industrielles de l'intelligence artificielle;
- au centre de l'axe se retrouvent les autres thèmes d'application de l'intelligence artificielle dans la prise de décision, la construction navale (thème : "conception assistée"), l'informatique biomédicale, l'imagerie et la reconnaissance des forme, la recherche documentaire
- vers le bas, les thèmes théoriques ("raisonnement", "apprentissage", " méthodologie").

Cette carte permet de voir comment s'organisent d'un point de vue thématique les références de ce corpus portant sur "l'intelligence artificielle".

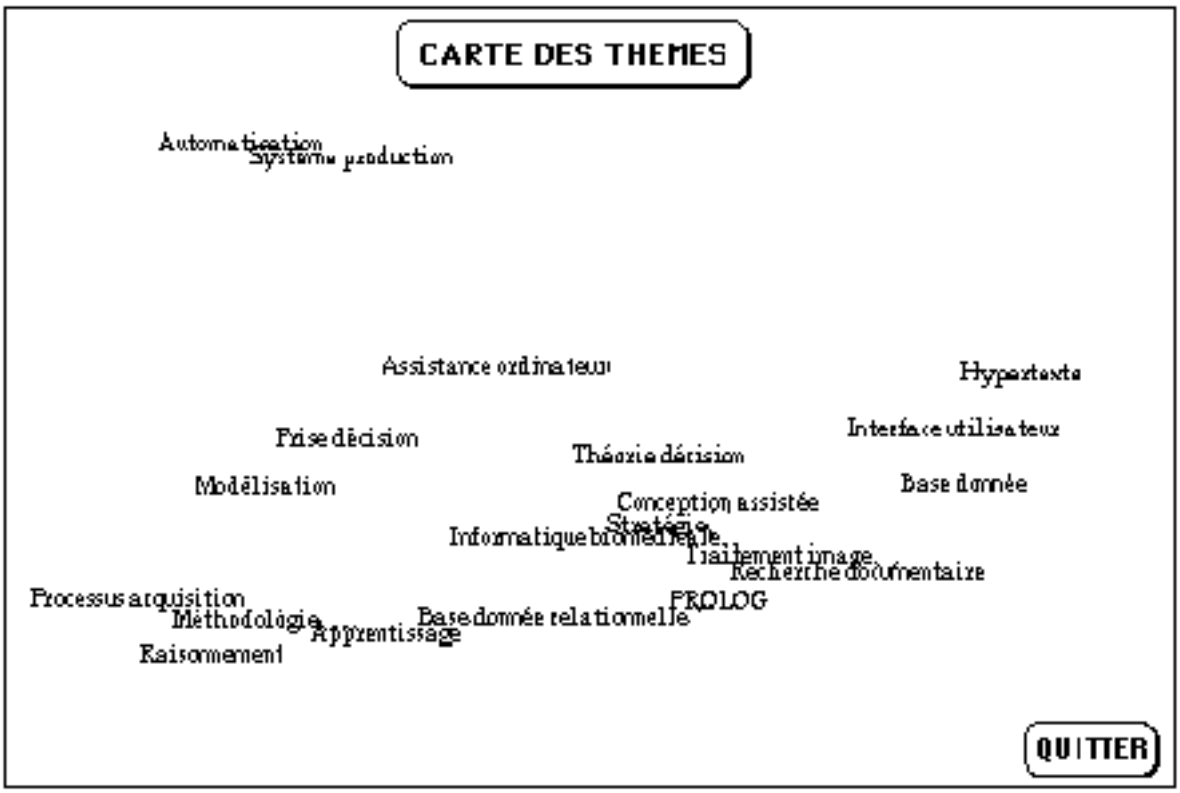


Figure 8 : Exemple de carte des thèmes obtenue avec NEURODOC

3.2.3 - Comparaison des cartes obtenues avec SDOC ou NEURODOC dans l'exemple proposé

Dans les deux cartes, on trouve 80% de thèmes communs, qui peuvent avoir des intitulés différents (40% de noms identiques), les thèmes applicatifs étant situés vers le haut, les thèmes théoriques étant plutôt situés vers le bas. Le fait que l'opposition thèmes applicatifs/ thèmes théoriques soit mise en évidence et rendue de la même manière sur les 2 cartes est fortuit. Dans le cas de NEURODOC, cette position s'explique par le contenu des thèmes. Dans le cas de SDOC, la position des thèmes est expliquée par leur structure. Ainsi, la position des thèmes applicatifs est due à la présence d'une indexation plus spécifique des documents associés. Ceci induit une forte cohésion au niveau des thèmes applicatifs. Les thèmes théoriques sont constitués de mots aux fréquences plus élevées. Leur cohésion est plus lâche, ce qui explique leur position vers le bas sur la carte SDOC.

Deux courts exemples pour illustrer les différences qui existent au niveau des cartes :
Prenons le thème Apprentissage obtenu avec NEURODOC. Il recouvre les thèmes Apprentissage et Parallélisme obtenus par SDOC. Sur la carte SDOC, ces deux thèmes sont très éloignés mais reliés par une liaison externe, rendant compte d'une liaison structurelle entre un thème spécifique Parallélisme, regroupant 7 documents, et un thème générique Apprentissage qui regroupe 19 documents, dont 4 appartenant au thème parallélisme.
A l'inverse, prenons le thème "système production" obtenu avec SDOC. Il recouvre les thèmes "automatisation" et "système production" obtenus par NEURODOC. La carte

NEURODOC montre que les thèmes “automatisation” et “système production” ont un contenu voisin et constituent un groupe de documents très spécifiques par rapport aux autres thèmes.

3.2.4 - Conduite d’une analyse avec SDOC ou NEURODOC

Bien que l’hypertexte facilite une démarche d’investigation par association d’idées, nous pensons que l’analyse de l’information peut être pilotée en usant d’indicateurs tels que ceux cités plus haut. Pour les deux méthodes, les étapes de l’analyse sont similaires :

- évaluer d’abord la qualité de la partition des mots-clés et des documents en sachant qu’il s’agit toujours de trouver un compromis entre une bonne lisibilité (un nombre de clusters pas trop élevé) et une moindre perte d’information ;
- étudier le tableau résumant les caractéristiques des clusters/thèmes, repérer les clusters/thèmes dominants/principaux et les clusters/thèmes secondaires ;
- étudier la ou les cartes des clusters/thèmes, puis le contenu des clusters/thèmes, afin d’appréhender l’organisation thématique du corpus documentaire.

Cette esquisse de méthodologie a pour unique ambition d’aider à l’exploration de résultats et ne devrait constituer en aucun cas un obstacle à l’intuition. C’est un moyen de disposer des premiers éléments constitutifs d’un dossier d’analyse sur lequel on peut s’appuyer pour étayer ses réflexions.

4 - Bilan et évolutions de la station de travail

Notre station de travail permet de caractériser et d’analyser par deux méthodes différentes un ensemble de références bibliographiques. Il nous semble important d’insister encore une fois sur la possibilité de mesurer les accords entre les résultats des deux méthodes d’analyse par des indicateurs globaux (réduction de donnée, taux de recouvrement, nombre de thèmes identiques ou voisins, taille des classes de documents). Il reste cependant que les méthodes offrent des représentations différentes : classes de mots-clés structurées par les relations de cooccurrences dans un cas, classes de mots-clés floues et recouvrantes représentées par des demi-axes dans l’autre. On a vu également que les cartes fournissaient des informations de natures différentes : indicateurs structurels et visualisation des réseaux locaux pour SDOC, oppositions des thèmes selon deux axes principaux pour NEURODOC. Cette richesse au niveau des représentations ainsi que la possibilité de comparer globalement les résultats justifient à notre avis la présence des deux méthodes au sein de la station, chaque méthode apportant un éclairage analytique particulier.

Les évolutions de notre station de travail à court, moyen et long terme :

- Amélioration de l’interface

L’interface actuelle pour le pilotage de la chaine de traitement infométrique est trop rudimentaire dans le cadre d’une utilisation occasionnelle de la station. Nous en avons fait l’expérience au cours de la formation d’un agent à nos outils. L’existence de générateurs d’interface MOTIF nous permet d’envisager avec confiance le développement d’une interface graphique pour le pilotage des modules de traitement et de visualisation. En effet, les fonctionnalités de la station de travail sont maintenant bien stabilisées.

- Amélioration des possibilités d'exploitation des résultats fournis SDOC et NEURODOC. Les prototypes que nous avons développés permettent à un utilisateur de visualiser la carte des thèmes, accéder à la description du thème (liste de mots-clés), puis d'accéder à la liste des titres (ou des auteurs ou des sources) des documents associés puis d'accéder à un document donné. A l'heure actuelle, l'utilisateur ne peut pas réellement poser de questions; il ne peut que naviguer par des chemins pré-établis. Pourtant, un responsable d'industrie désireux de connaître les sociétés ou les équipes de recherches qui travaillent sur les mêmes thèmes que son équipe ou de suivre les thèmes sur lesquels travaille une société concurrente, aura envie "d'interroger" la carte des thèmes par frappe au clavier d'une équation booléenne de mots-clés, par sélection d'un groupe de documents représentatifs du problème qu'il se pose, une liste d'auteurs, une date de publication, un ensemble de revues, des organismes d'affiliation. L'utilisateur devrait pouvoir exprimer des requètes complexes sur les thèmes mis en évidence par nos outils infométriques, effectuer des annotations, et stocker les requètes effectuées pour reprendre une analyse là où il l'avait laissée. Fournir ces fonctionnalités a fait partie dès le début de nos objectifs. Ainsi, dans son interface Hypercard actuelle, NEURODOC permet de sélectionner un mot-clé et de le situer sur la carte des thèmes par mise en gras des thèmes où figure ce mot-clé. Mais les temps de réponse sont tels qu'on ne peut l'envisager sur des corpus importants. Nous sommes donc à la recherche d'autres supports pour une telle réalisation. L'émergence d'une nouvelle génération de systèmes hypertextes sur l'internet nous permet d'envisager aujourd'hui ce développement avec plus d'optimisme.

- Intégration d'autres techniques d'analyse et de visualisation des résultats. Considérant que l'INIST constitue un observatoire privilégié des sciences, nous désirons appliquer toute méthode pertinente pour cette observation. Nous pensons que le soin que nous avons porté à la conception de cette station (notamment au niveau de sa modularité) facilitera ce type d'intégration. Notre ambition n'est pas de vouloir redévelopper des techniques d'analyse existantes, mais plutôt d'être capable d'intégrer leurs résultats facilement. La station jouera alors un rôle d'intégrateur en tant que moyen de consultation.

Remerciements
La station d'analyse infométrique est le produit d'une équipe. Nous remercions nos collègues du Programme de Recherche en Infométrie, Xavier Polanco, Dominique Besagni, Chantal Muller et Jean Royauté pour leurs développements, critiques et réflexions ainsi qu'Alain Lelu pour ses apports (écrits et verbaux) concernant la symétrie des deux méthodes.

Nota
Notre bibliographie est volontairement circonscrite à notre filière méthodologique dans la mesure où notre objectif dans cet article n'est pas de comparer notre station de travail ou les méthodes utilisées avec d'autres, mais de présenter une réalisation du programme de recherche infométrie, et une démarche d'analyse. Diverses études ont été menées à partir des outils présents sur cette station : étude TELETHESE "Santé, Sciences et Sciences Sociales" (40 000 thèses analysées en mars 1992 pour le ministère de l'éducation nationale), dans le domaine des cognisciences [DUCLOY et POLANCO 1992],

l'économie de l'information [POLANCO et al 1993b],.la sociologie (14 000 références de la base FRANCIS en sociologie de 1989 à 1991) [POLANCO et GRIVEL 1994], l'histoire sociale allemande à partir de la base SOLIS de l' InformationsZentrum Sozialwissenschaften (IZ) [GRIVEL et al.1995], la revue *Scientometrics* [POLANCO et FRANCOIS 1994], etc.

5 Références

CALLON M., COURTIAL J-P., TURNER W.A., BAUIN S. 1983 - "From Translation to Problematic Networks: An Introduction to Co-Word Analysis" in Social Science Information, vol. 22, pp. 191-235.

CALLON M., LAW J., RIP (eds). 1986 - "Mapping the Dynamics of Science and Technology" LONDON: The Macmillan Press Ltd.

CALLON M., COURTIAL J-P., PENAN H.1993 - "La scientométrie" - Presses Universitaires de France, collection "Que sais-je", Paris.

COURTIAL J-P. 1990 - "Introduction à la scientométrie", Anthropos - Economica, Paris.

DUCLOY J., CHARPENTIER P., FRANCOIS C., GRIVEL L. 1991 - "Une boite à outils pour le traitement de l'information scientifique et technique", Génie logiciel et systèmes experts, n°25, pp 80-90, Paris.

DUCLOY J., POLANCO X.1992 -"D'une boite à outils à la description du domaine des cognisciences", Journées d'étude ADEST "Prendre la mesure des sciences et techniques : la scientométrie en action", Paris 1-11 juin 1992.

GRIVEL L., LAMIREL J.C. 1993 - "An analysis tool for scientometric studies integrated in an hypermedia environment", ICO93, 4th International Conference on Cognitive and Computer Sciences for Organizations, Montreal, (Quebec) Canada, pp146-154, 4-7 mai 1993.

GRIVEL L., MUTSCHKE P., POLANCO X. "Thematic mapping on bibliographic databases by cluster analysis: a description of SDOC environment with SOLIS", à paraître

LEBART L., SALEM A. 1988 - "Analyse statistique des données textuelles", DUNOD, Paris 1988, 207 pages.

LELU A. 1990 - "Modèles neuronaux pour données textuelles - Vers l'analyse dynamique des données" - Journées ASU de statistiques, Tours, France.

LELU A. 1990 - "Modèles neuronaux de projection associative at analyse des données" - Approches symboliques et numériques pour l'apprentissage de connaissances à partir des données - sous la direction d'E. DIDAY et Y. KODRATOFF, pp 283-305, CEPADUES, Toulouse.

LELU A. 1993 - "Modèles neuronaux pour l'analyse de données documentaires et textuelles" Thèse de doctorat de l'université de Paris VI.4 mars 1993, 238 pages.

LELU A. et FRANCOIS C. 1992 - "Automatic generation of hypertext links in information retrieval systems", communication au colloque ECHT'92, Milan, D. Lucarella & al. eds, ACM Press, New York.

PETERS H.P.F., VAN RAAN A.F.J. 1993 - "Co-word based science maps of chemical engineering, Part II : Representations by combined clustering and multidimensional scaling" Research Policy, vol.22, 1993, p.47-70.

POLANCO X. 1995 "Aux sources de la scientométrie", *SOLARIS* n° 2, Presses Universitaires de Rennes, p. 13-79 ; http//www.info.unicaen/bnum/jelec/Solaris.

POLANCO X. et FRANCOIS C. 1994 - “Les enjeux de l’information scientifique et technique à travers une analyse d’infométrie cognitive utilisant une méthode de classification automatique et de représentation conceptuelle (NEURODOC)”, Actes du colloque ORSTOM/UNESCO “Les sciences hors occident au XXè siècle, Paris 19-23 septembre 1994.
POLANCO X. et GRIVEL L. 1995 - “Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United State of America), Internation journal of Scientometrics and Informetrics, Vol1 (2),pp123-137
POLANCO X., FRANCOIS C., BESAGNI D., MULLER C., GRIVEL L 1993a - “Le programme de recherche infométrie”, Les systèmes d’information élaborée, Ile Rousse 9-11 juin 1993
POLANCO X., FRANCOIS C., BESAGNI D., MULLER C., GRIVEL L 1993b - “Un exemple de traitement de l’information par une approche infométrique : le cas de l’économie de l’information”, 3ème conférence internationale sur la recherche en informations - Nouvelles technologies de l’information : les défis pour la recherche en économie de l’information, Poigny la Forêt, France, 11-13 juillet 1993.
TURNER W. 1994 - “Penser l’entrelacement de l’Humain et du Technique : les réseaux hybrides d’intelligence “- Solaris n°1 “Pour une nouvelle économie du savoir”, Presses universitaires de Rennes, p.21-50.

Chapitre 6[1]

Assister l'analyse de l'IST par la génération automatique d'hypertextes dynamiques à l'ère d'internet et du World Wide Web.

Ce chapitre décrit les choix de conception d'un générateur d'applications hypertextes adaptés à la démarche d'analyse précisée dans le chapitre précédent :

1. S'appuyer sur un Système de Gestion de Bases de Données (SGBD) et sur une modélisation relationnelle des données infométriques

Ce système gére plusieurs types de documents, les clusters et les données structurées à analyser (qui peuvent d'ailleurs être de différents types). L'idée est de modéliser les données structurées et les clusters obtenus à partir de ces données, de tel façon que la plupart des opérations d'analyse puissent être exécutées par des requêtes SQL (Structured Query Langage. Ainsi, un couplage des technologies Base de Données et Hypertextes donne la possibilité de mettre en relation tout élément constitutif d'une information structurée (référence bibliographique, brevet) avec les thèmes (clusters) obtenus par classification automatique. Dans cet article, nous avons pris comme exemple les résultats du programme NEURODOC sur des données bibliographiques.

2. Le choix du système hypertexte.

Celui-ci doit pouvoir communiquer avec le SGBD. Le World Wide Web (WWW) répond à ce besoin. Ce système hypertexte distribué, peut facilement être étendu pour communiquer avec les SGBD. Les avantages d'une passerelle WWW-SGBD sont énormes par rapport à une structure arborescente de documents hypertextes textuels (même construite automatiquement). L'administration du site des données et des utilisateurs est facilitée car les liens entre documents sont calculés dynamiquement et n'ont pas à être maintenus. De plus, un simple export de la base suffit à préserver le site. Un bon niveau de confidentialité peut être garanti car les autorisations d'accès peuvent être gérées au niveau du serveur WWW et du SGBD.

3. Le choix de SGML.

En exploitant la dualité existant entre structure d'arbre SGML et schéma E/A (Entité/Association), HENOCH assure deux fonctions principales: alimenter le SGBD à partir de tout type de document structuré conforme à la norme SGML et établir une interface WWW-SGBD.

4. L'interface utilisateur.

L'interface utilisateur propose deux types de navigations complémentaires : une exploration intuitive basée sur la métaphore de la carte, et un mode de recherche basé sur la métaphore "Qui fait Quoi, Où, avec Qui, Quand, dans quelles sources (revue, congrès, ...)". Dans les deux cas, la navigation est assurée par l'exécution de requêtes SQL sur la base de données infométriques.

---

[1] Grivel L., Polanco X., Kaplan A. 'A computer System for Big Scientometrics at the Age of the World Wide Web', Scientometrics, vol.40, N°3, 1997, 493-506, 1997, et in proceedings of the 6th International Conference on Scientometrics and Informetrics, Jerusalem, 131-142, 1997.

1. INTRODUCTION

This paper stresses the "computerized framework" that informetrics need to develop their industrial dimension. If we consider the two last international conferences on informetrics (Berlin 1993 and Chicago 1995), the computer point of view has been relatively neglected by the informetric community which seems to be a community of users which is not concerned with the creation of computer means. At least, that is what appears if we compare last conferences on Information Retrieval (SIGIR) with Informetrics conferences.

We argue that an informetric method should not only be characterized in terms of its mathematical representational adequacy, but also in terms of its computational architecture and effectiveness. A computationally effective informetric system should explain the relationships between the nature of statistical representation, the effectiveness of techniques, and the computational architecture in which the computations/informetric techniques are performed. Cluster analysis and map-based representation formulation are examples of such informetric techniques.

The INIST's Informetric Research Program (in french "Le Programme de Recherche Infometrie (PRI)") is at the origin of a global informetric system for the analysis of scientific and technical information (STI) (Polanco 1996). This system or computational architecture uses :

1. Computational linguistic programs which provide mechanisms of terminological extraction on full text (in English and in French) in order to replace manual indexing and to build more complex linguistic knowledge indicators than simple keywords (Polanco & al. 1995).
2. Clustering and mapping programs such as NEURODOC (Polanco & François 1997) and SDOC (Grivel & Polanco 1995; Grivel & al. 1995).
3. HENOCH system.

HENOCH system organizes the results of NEURODOC or SDOC in a relational database management system (RDBMS), and provides them to users through a client/server architecture based on World Wide Web (WWW) via Internet or Intranet.

HENOCH development started in September 1994, with a joint project with ESIAL (Ecole Supérieure d'Informatique et Applications de Lorraine) implemented by a group of software engineer-students during their last year of study. In march 1995, a mock-up proved the feasibility of the system with a freeware RDBMS (Requiem). Six months later, a prototype was built by one of these students on a SUN workstation with Oracle DBMS.

HENOCH is now fully operational. It is an element of the INIST information system since its transfer on an HP mini-computer. Beta-testing by INIST partners/customers has started in summer 1996 and will end in March 1997. This computer system is currently used in many applications in the industrial sector as well as in research. HENOCH was presented at the central office of the CNRS the 25th of June 1996.

In section 2, we will explain our choices in terms of information and hypertext systems. In section 3, we will present the software components of HENOCH which is used as a

generic environment for storing SGML documents into a relational database, and to make this data accessible via the Web. This is exemplified in the particular case where the stored data are clusters and bibliographical data. In section 4, we will show the benefits of this environment for information analysts if the data is properly modeled for information analysis.

2. HENOCH SYSTEM

The information analysis process in the frame of scientific watch is a mix of informal exploration and of specific requests like "Who does What and Where, When, ...". Hypertext technology extended with retrieval techniques (coming from documentary systems or from DBMSs) address this need (Balpe & al. 1996) [1]. We will now explain the interest of extending our informetric processing chain with a database system coupled with an hypertext system.

2.1 Database system

Information analysis operates here in an "informetric processing chain" relying on SGML (Standard Generalized markup Language) (Goldfard 1990 [2]; Herwijnen 1990 [3]). An example of bibliographic data description with a tagging based on SGML is given in section 3.1. The markups (or tags) only describe the logical structure of the documents. Thus, it is very easy to associate procedures or treatments to tags for a given application. This association mechanism enables several applications (formatting, linguistic, clustering, hypertext generation tools, ...) to work on the same document description. SGML is particularly convenient for automatic information processing. In our informetric chain, all intermediary results are stored in a hierarchical SGML files system (Ducloy & al. 1991) [4]. With this technology, combining clusters (data on topics) and bibliographic data requires customized programs to compute all necessary combinations before being able to generate the corresponding hypertexts. Even if this programming step may be facilitated by the use of SGML based tools, this is a repetitive task. According to Small (1995), "Existing bibliographic search software simply does not allow [combination of bibliographic data elements]. Since relational databases[5] are designed explicitly to relate data elements to one another, they would seem a natural choice for bibliometric analyses".

The idea is to model bibliographic data and clusters so that most operations ("Who does What and Where, When, ...") are undertaken with SQL statements. RDBMSs considerably facilitate data and user administration, because these solutions are now mature and reliable. Although the relational model[5] has its drawbacks compared to object-oriented models[6], one may notice that RDBMSs tend to become hybrid systems by merging relational and object features, thereby becoming more adapted to the management of structured textual data.

2.2 Hypertext system

Classical hypertext systems such as Winhelp or Hypercard essentially allow static navigation, possibly enhanced by some keyword interrogation facilities (Winhelp) or script language capabilities (Hypercard). We have tested these two systems. This kind of hypertext systems are adequate to publish low-cost documentary products on CD-ROMs

or floppy disks, but they cannot be used for cooperative work or easily extended. They do not provide the navigation mechanisms (dynamically computed nodes and links) needed within a dynamic information system. They are not available on every hardware platform. The World Wide Web (WWW) is much more open and extensible. WWW is a distributed hypermedia system on the Internet, organized as a client/server architecture. WWW clients are available for virtually any hardware platform. WWW can be easily extended (by means of plug-ins, java classes, ActiveX, CGI programs) to implement both exploration and interrogation facilities or cooperative work.

Consequently, we decided to develop a RDBMS-WWW gateway which lets the user access informetric databases from his favorite WEB browser. So that the information analyst will be insured to get up-to-date information with a user-friendly interface on a basic PC or Macintosh.

Web designers will notice that a RDBMS-WWW gateway greatly facilitates the administration of a WWW server by avoiding the tree-structure of links usually maintained on most sites. Considering security, a simple database export is sufficient to preserve the whole site. A high degree of confidentiality can be obtained because access authorizations can be managed both within the WWW server and the database server.

All of these features are important when many customers are expected to access informetric analyses. Being designed to easily store any SGML document into a relational database, and to make these data accessible via WWW, HENOCH meets all these underlying requirements.

3. HENOCH SOFTWARE CHARACTERISTICS: A GENERIC ENVIRONMENT

HENOCH is made of three C++ programs (Skelettor, Convertor and ICGI[7]). Overall, Skelettor and Convertor feed in data into the RDBMS (figure 1) and ICGI is the WWW-RDBMS gateway (figure 2). From a software engineering point of view, HENOCH meets two requirements, i.e. to integrate itself into our informetric system and to provide for the reusability of its components for other applications (based on SGML, RDBMS, WWW). HENOCH is designed as an applications generator. HENOCH components permit to supply any RDBMS with SGML documents. Then, the WWW-RDBMS gateway is as generic and extensible as possible (use of templates of HTML pages containing both classical HTML tags and embedded SQL calls, advanced functions of presentation, such as graphs, histograms, maps).

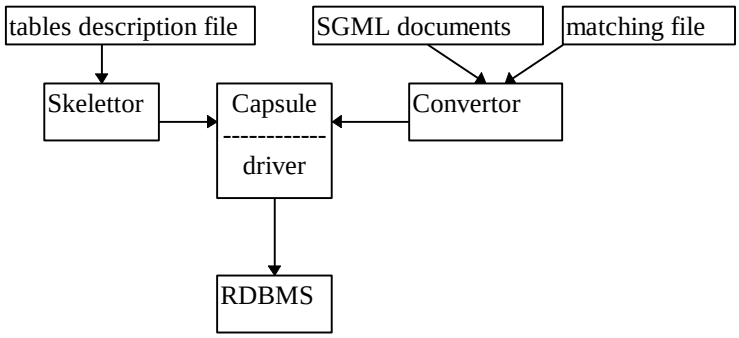


Figure 1 : The RDBMS supplying process

3.1 Conversion of SGML documents into database tables

As shown in figure 1, Skelettor creates the "skeleton" of the database i.e. the basic structures to represent the data. Skelettor takes as an input the tables description and the code of the insertion procedure (in PLSQL, procedural SQL) associated to each table. Then, Convertor reads the SGML data files and stores the data in the database. These two C++ programs communicate with the RDBMS through an API (Application Programming Interface), an abstraction level which makes transparent the calls to the specific functions of the chosen DBMS. The API called Capsule encapsulates the creation and manipulation functions of the database system. By using the API Capsule, the programs are independent from the RDBMS.

To obtain independence from data structures, Convertor uses a matching file between SGML data elements and the tables used: The main idea is to handle the SGML documents as trees. The tree model allows random access to any node in the tree at any moment, so that solves forward references.

Below is a model of a matching file: We call SGMLpath a method of designating a particular node in the tree. A data contained in a node SGMLpath#1 is stored in a variable V#1 which is a parameter of the insertion procedure Proc#1.

```
TABLE_NAME:
V#1    SGMLpath#1
query :
begin
/* the insertion procedure to execute */
Proc#1(:{V#1})
end;
```

The following is an example of a bibliographic data description in SGML:

```
<record>
<NO>90-0128293</NO>
<TI>Density-dependent interactions between seedlings of Dactyloriza majalis
(Orchidaceae) in symbiotic in vitro culture</TI>
<AU>RASMUSSEN (H.);JOHANSEN (B.);ANDERSEN (T. F.)</AU>
<AF>
        <NA>Univ. Copenhagen, botanical lab.</NA>
        <TO>Copenhagen 1123</TO>
        <CO>DNK</CO>
</AF>
<DT>Publication en serie</DT>
</record>
```

This is the matching file corresponding to this type of document:

```
TABLE AFFILIATION:
```

```
Name        record/AF/NA
Town        record/AF/TO
Country             record/AF/CO
query :
begin
/* the insertion procedure to execute */
INS_AFFILIATION(:{NAME}, :{TOWN}, :{COUNTRY})
end;
```

This matching file is used by Convertor to identify the SGMLpaths needed to extract the data. Convertor parses the document, searches for all these paths and stores them in an associative array (variable<-->data). The instantiated insertion procedures are then executed.

3.2 A generic and extensible WWW-RDBMS gateway

Figure 2 shows a WWW server triggering a WWW-RDBMS gateway called ICGI, a program compliant with the Common Gateway Interface CGI, protocol of communication between an external program and a Web server. ICGI has been designed as a C++ object class whose main functions include: [a] the parsing of its arguments, and especially the one which specifies the type of graphical display to be built, and according to this type of display, [b] the transmission of its arguments to the involved object class used to interpret the other parameters (like for instance, HTML template containing some SQL queries, user name and password, size of the map to build, ...).

Using these parameters, the involved object class creates the DBMS connection via the Capsule module we have previously described, sends SQL queries to the RDBMS kernel (always via the Capsule), formats on the fly into HTML the rows returned by the database, and lastly disconnects from the DBMS.

To achieve new advanced functions of presentation or some complex SQL statements using intermediary results, ICGI can be extended by creating a new specialized sub-class.

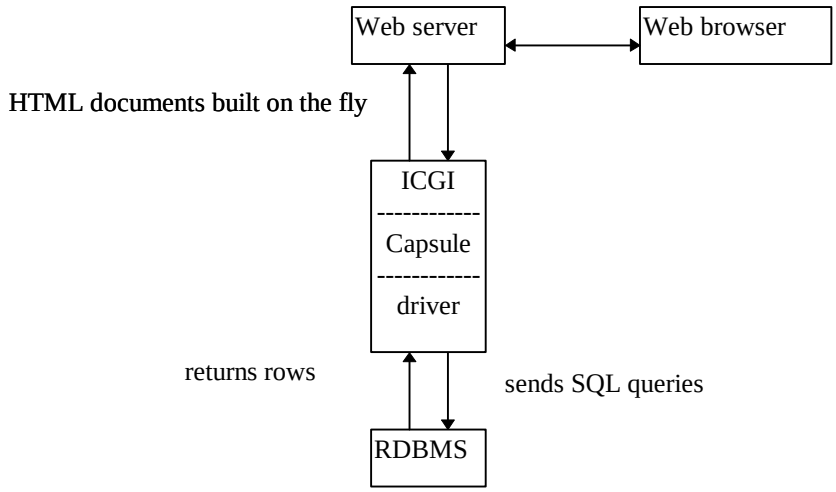


Figure 2 : ICGI, a WWW-RDBMS gateway

3.3 About HENOCH software components

Convertor uses a SGML toolkit called ILIB[4] which is able to work both on character streams and on trees. Another possibility about the conversion of SGML documents into database tables is to use a public SGML parser tool kit[2] integrating an API. Consequently, a program using this API is able to trigger actions in regard of a context in order to generate an input file for a data loading product such as SQL*Loader (Oracle).

WWW-RDBMS gateways were at their very beginning in 1995. The unique way to trigger or run an external application (such as a RDBMS) from a Web server was to write a CGI program. Unfortunatly the CGI, based on HTTP protocol, is not a very efficient mechanism and does not integrate the notion of transaction. Consequently, it generates a great number of request and overloads the network. Next version of HENOCH [9] will use Java applets based on JDBC [8] to solve this kind problem.

HENOCH, integrated in our SGML-based informetric chain, is an effective rapid prototyping environment which allows to test and validate new functionalities very quickly in real size at a very low cost. The next section illustrates an information analysis environment generated by HENOCH from SGML informetric data format.

4. AN EXAMPLE OF INFORMATION ANALYSIS ENVIRONMENT

4.1 Relational modeling of informetric data

We call informetric data the results of clustering and mapping programs applied onto a corpus of bibliographic data in a particular scientific field. The relational data model depends on the features of these results. Here, we take the results of the NEURODOC program as an example. The data model slightly differs in the case of citation, co-citation [9] or coword analysis.

The two main components of NEURODOC are: [a] Cluster analysis which groups the documents by cluster, and therefore also the authors, their affiliations and the journals in which they were published. This cluster analysis is achieved using the axial k-means method. [b] A factor representation of topics (or clusters) identified above based on the principal component analysis (PCA). The keywords are used indicators of the knowledge content of documents. (Polanco & al. 1997).

A NEURODOC cluster consists of a ranked list of weighted keywords and a ranked list of weighted documents. A label is attached to each cluster. A cluster has coordinate values on a bi-dimensional map. Each bibliographic reference is composed of fields (possibly in several languages), such as title, abstract, authors, affiliations, publication date, document type, etc.

These two SGML document types (clusters and bibliographic data) are considered as two composite entities and are broken down into several interrelated tables. (cluster table, cluster-keyword table, document table, author table, keyword table, affiliation table, ...). The document Id and the cluster Id play a key role. Once the bibliographic data is broken down into a set of interrelated tables, the document Id is used to relate the

tables back together, and so does the cluster Id. To relate clusters and bibliographic data, the document Id and the cluster Id are used together.

4.2 Hypertext interface

The interface is derived from our previous work (Grivel & François 1995 a, 1995 b) [10] and takes advantage of the relational model and of user comments. It is clear that the definition of a user interface requires compliance with some principles or guidelines which will not be described here as it is not within the scope of the paper.

The informetric database interface provides two types of navigation, which are complementary to analyze information: a) an intuitive exploration mode based on the map metaphor and b) an assisted searching mode based on the "Who does What, and Where, with Whom" metaphor.

The screenshots come from a study on industrial enzymes by Harry Rothman (Director, Centre for Science & Technology Policy, University of the West of England, Bristol), based on data extracted from PASCAL INIST's database. The clustering and mapping application is NEURODOC.

4.2.1 Exploration mode

The information analyst can use navigation possibilities which correspond to predefined requests on the database containing the informetric data. By simple clicking on a link, the user can build a cluster map (Figure 3) for having an idea of what the information space looks like. By clicking on a cluster name (here 'Enzyme inhibitor') in this map, he can zoom its description by a weighted keywords list (Figure 4). He can then examine its related list of document titles (Figure 5) (or authors or sources), and select one of these titles to access its full bibliographic description (Figure 6). He can also use the keywords composing the cluster to access the bibliographic references belonging to this cluster and indexed with these keywords.

Map-based navigation helps to make a global analysis of the information landscape for a given subject. It is also an invaluable aid for a user to first explore a domain which is at the outer edges of its usual area of interest. But it is insufficient to answer to a question like "Who does What and Where, When, with Whom ?".

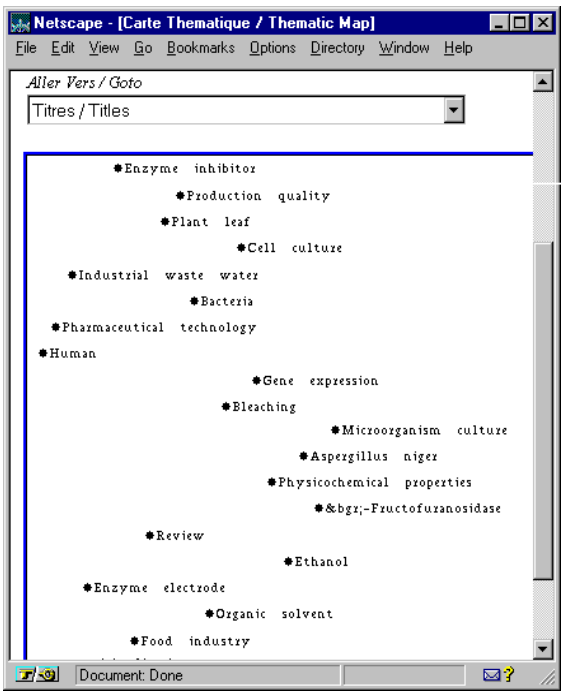


Figure 3 : clusters map

Figure 4 : cluster 'Enzyme inhibitor'

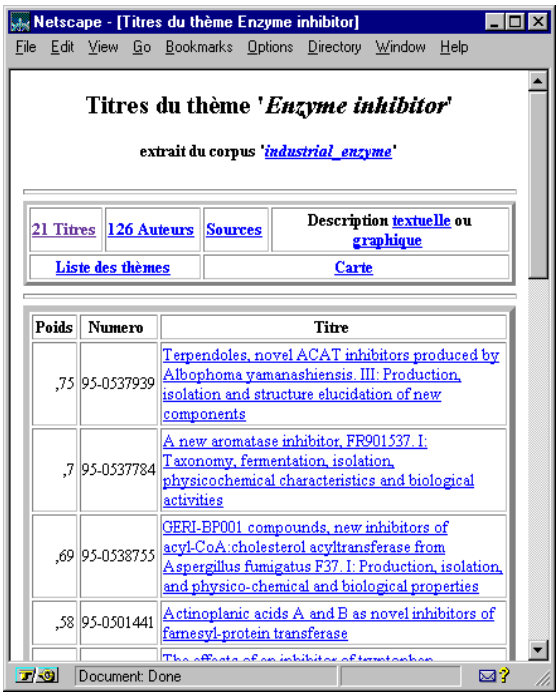


Figure 5 : titles related to 'Enzyme inhibitor'

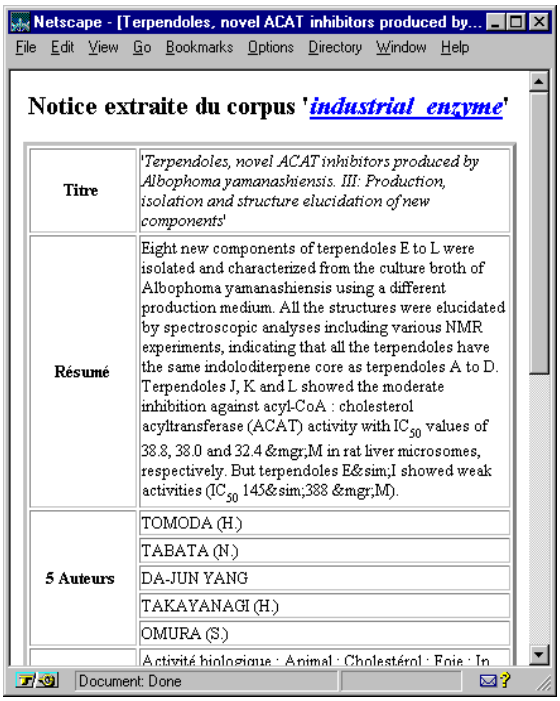


Figure 6 : bibliographic data

4.2.2 Assisted searching mode

Our goal is not information retrieval by clustering analysis [11] but information analysis through a user-friendly interface specifically adapted to the execution of this function by a user agent. This is what we mean by assisted searching mode based on the "Who does What and Where, When, with Whom ?" metaphor. The user can search by authors names, affiliations, keywords, journals titles (or other information sources) in order to know in which clusters these elements are. For obtaining a global idea of the areas of

interest of a company, he can express the following queries by a simple navigation path (Figures 7,8): select all the affiliations of the corpus; select all the documents whose affiliations begin by "ARS"; select all the clusters related to documents whose affiliations begin by "ARS" and count the number of documents for each cluster.

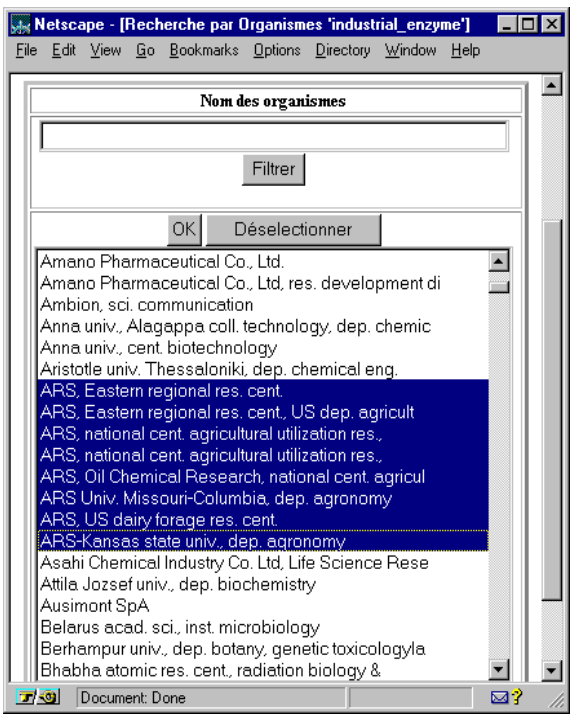


Figure 7 : search by affiliations

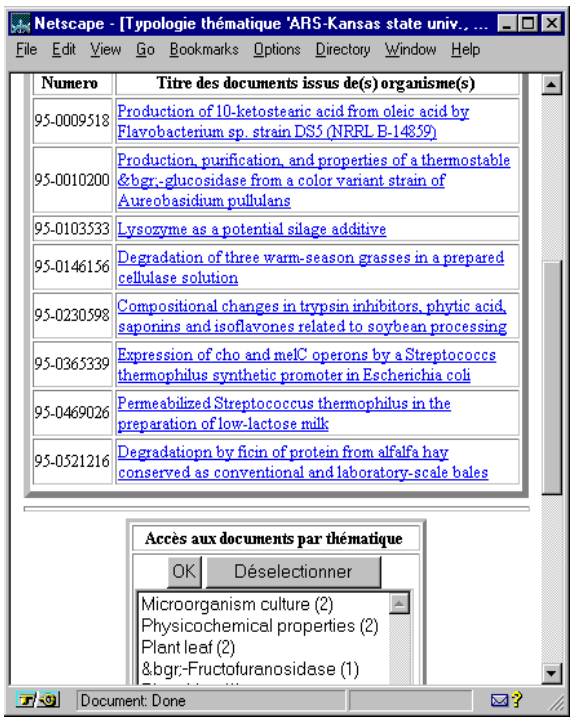


Figure 8 : results of the search by affiliations

This WWW-RDBMS based interface permits to build complex queries very easily without requiring knowledge in SQL query language. It provides the information analyst with a simple and efficient means to intersect or join some items of information featuring a scientific or technical activity sector.

5. CONCLUSION

In this paper, we have proposed a computer system for "Big Scientometrics" at the age of the World Wide Web. This computer system is a generator of informetric databases. HENOCH is a generic environment to store in a relational database any SGML document produced by an informetric environment [12] and to make these data accessible via the Web. "Big Scientometrics" requires a significant computer environment with computational linguistic techniques, statistical methods, graphic tools and an efficient storage and management system. We think that the technical architecture proposed here can be applied to other informetric environments. Application to other document types than bibliographical data or clusters is straightforward. It corresponds to a mapping between SGML tree structure and a relational model.

Today, HENOCH contains around 20 bibliographical data corpora on different subjects in various formats. Each corpus is the result of a request on a database (PASCAL, FRANCIS, SCI, ...) used as an information documentary profile on a given subject (for instance human ressources, natural energies, linguistic ingeneering and natural language

processing, information technology, artificial intelligence and expert systems, etc). Each profile has been clustered, mapped by NEURODOC or SDOC, and stored in the informetric database for beta-testing under WWW by our partners until the end of March 1997.

ACKNOWLEDGMENTS : We would like to thank C. Broussaudier, B. Levy, who, with A. Kaplan, were members of the ESIAL team working on the first phase of this project. Special credit should also be paid to Mrs Brigitte Jaray, professor at ESIAL, for her guidance in this initial phase.

6. REFERENCES

BALPE J.P, LELU A., PAPY F., SALEH I. (1996)
Techniques avancées pour l'hypertexte, Paris, Editions Hermès.

CODD E. F.(1970)
A relational model of data for large shared data banks, *Comm. of the ACM*, Vol13 (6): 377-387.

DUCLOY J., GRIVEL L., LAMIREL J.C., POLANCO X., SCHMITT L. (1991)
INIST's Experience in Hyper-Document Building from Bibliographic Databases. *Proceedings of Conférence RIAO 91*, Barcelone (Spain), vol 1.

GOLDFARB C.(1990)
The SGML Handbook, Oxford, Oxford University Press.

GRIVEL L., MUTSCHKE P., POLANCO X. (1995)
Thematic mapping on bibliographic databases by cluster analysis: a description of the SDOC environment with SOLIS, *Journal of Knowledge Organization*, vol. 22, (2): 70-77.

GRIVEL L., FRANÇOIS C (1995a)
Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique, *SOLARIS*, n° 2, Presses Universitaires de Rennes: 81-113., and also on internet, http://www.info.unicaen.fr/bnum/jelec/Solaris

GRIVEL L., FRANÇOIS C (1995b)
Conception et développement d'un système d'information dédié à la veille scientifique, basé sur les sorties des outils de classification thématique : SDOC et NEURODOC , In : BALPE J.P, LELU A., SALEH I,.Eds, *Hypertexte et hypermedia, réalisations, outils et méthodes*, Paris, Editions Hermès: 109-118.

HERWIJNEN E. (1990)
Practical SGML, Kluwer Academic Publishers.

POLANCO X., GRIVEL L (1995)
Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United

State of America), *International Journal of Scientometrics and Informetrics*, Vol.1, (2): 123-137.

POLANCO X., GRIVEL L., ROYAUTE J. (1995)
How to do things with terms in informetrics: terminological variation and stabilization as science watch indicators, In: Michael E.D Koenig, Abraham Bookstein (Eds), *5th International Conference of the International Society for Scientometrics and Informetrics*, Learned Information Inc. Medford NJ: 435-444.

POLANCO X. (1996)
La notion d'analyse de l'information dans le domaine de l'information scientifique et technique, *Conference INRA-Information scientifique et technique*, 21-23 october, Tours, France (forthcoming).

POLANCO X., FRANÇOIS C., KEIM J.P. (1997)
Artificial Neural Network Technology for the classification and Cartography of Scientific and Technical Information, to be published in Proceedings *6th International Conference of the International Society for Scientometrics and Informetrics*, Jerusalem, June 16-19.

SALTON G (1989)
Automatic Text Processing: The transformation, Analysis and Retrieval of Information by Computer. Reading, MA:Addison-Wesley

SMALL H. (1995)
Relational bibliometrics, In: Michael E.D Koenig, Abraham Bookstein (Eds), *5th International Conference of the International Society for Scientometrics and Informetrics*, Learned Information Inc. Medford NJ: 525-530.

ZITT M. , BASSECOULARD E.(1994)
Development of a method for detection and trend analysis of research fronts built lexical or cocitation analysis, *Scientometrics*, Vol.30, (1): 333-351.

ZITT M. , BASSECOULARD E. (1996)
Reassessment of co-citation methods for science indicators: effects of methods improving recall rates, *Scientometrics*, Vol.37, (2): 223-244.

7. NOTES

1. We have found in this book the definition of a general computer framework to generate hypertexts. Although our informetric platform had been designed for another purpose (information analysis), many ideas about hypertext generation expressed in this book are implemented in our global informetric system.

2. Dr Charles F. Goldfarb (IBM) is the inventor of the SGML language (http://www.sil.org/sgml/sgml.html). SGML is a meta language used to build specific markup languages. The best known markup language based on SGML is HTML (HyperText Markup Language), which describes the logical representation of Hyperdocuments on the World Wide Web. A markup language based on SGML can be built for each class (or type) of documents. The SGML Handbook is a reference

for those who want to understand in very detail the SGML standard in order to developp SGML-based tools like for instance parsers (http://www.sil.org/sgml/publicSW.html#parserTools).

3. The book of Dr Eric van Herwijnen can be given both for beginners to use SGML and for programmers to implement SGML. It is really practical !

4. A SGML toolkit called ILIB (Ducloy & al. 1991) was developped at INIST from 1990 to 1993. SGML is used to describe data (for instance bibliographical data) whatever their source,and also intermediate data between programs communicating by pipe. There is also an API which permits to manipulate SGML documents as a tree.

5. Codd is the inventor of the relational model. In this model, data are represented by tables. Basically, a table can represent either an entity or a relationship between entities. A table is comprised of rows and columns. Each column of a table represents one attribute of an entity. Each row represents one occurrence of an entity or relationship represented in a table. The table manipulations are insured by a set of algebraic and relational operations (Cartesian product, union, projection, selection) in SQL (Structured Query Language), a normalized language to interrogate a RDBMS.

6. The main concept of the object oriented approach is encapsulation. Data and treatments are integrated in a same entity: the object. A class describes a family of objects of same structure and behaviour. The notion of generalization/specialization permits to describe inheritance relations between classes. Other mechanisms (not directly linked to the object model) may exist. For instance the composition: an object can be composed of objects. The object model is an answer to some drawbacks of the relational model; in particular its unability to completely describe the semantic of complex structures by relationships between entities.

7. The name of our functions or of our programs often begin by I to mean Inist. ICGI means Inist Common Gateway Interface. ILIB means Inist LIBrary.

8. Java is an object-oriented programming language and environment developed by Sun Microsystems. Java programs (called applets) can be included in HTML pages and be run on a Web browser. The Java platform is continuously enriched by various APIs. For instance, JDBC API (Java Database Connectivity) provides the means to connect to any RDBMS and to embed SQL statements into Java applets. (htpp://java.sun.com)

9. In the frame of a common project with Michel Zitt (Zitt & Bassecoulard 1994, 1996), we are currently developping a Java interface to diffuse co-citation analysis results on INRA intranet. Based on JDBC, the developed programs will be used both in HENOCH (under Oracle) and in SAS environment. Special attention will be paid on graphical outputs, taking advantage of the Advanced Window Toolkit (AWT), a set of classes and interfaces classes for building sophisticated graphical interfaces.

10. In the first article, we compare SDOC and NEURODOC and suggest scenarii to analyse and qualify theirs results. A primary hypertext interface is demonstrated based

on these scenarii. The second article can be considered as the "birth certificate" of HENOCH as an interface to analyse SDOC and NEURODOC results. HENOCH is specified from a functional point of view by taking into account some drawbacks noticed in the primary interface.

11. In section 10.2 "Automatic Document Classification" (Salton 89), G. Salton shows how clustering analysis can be used in information retrieval for both searching and browsing a collection of documents. In this case, “the clustered file provides efficient file access by limiting the search to those document clusters which appear to be most similar to the corresponding queries”. In our case, we use clustering methods for information analysis. In this aspect, assisted searching based on the "Who does What and Where, When, with Whom ?" question is not only browsing or searching information but the dynamic calculus of strategic indicators (for instance authors or countries productivity or centers of interest, ...)

12. The informetric platform is composed of a natural language processing environment (in French and in English) called ILC platform (Polanco & al. 1995), NEURODOC (Polanco & François 1997) and SDOC (Polanco & Grivel 1995). Recently, it has been used in collaboration with INRIA (Institut National de Recherche en Informatique et en Automatique) to experiment knowledge acquisition and structuration from corpora on the field of agriculture. "Acquisition et structuration des connaissance en corpus: éléments méthodologiques", Muller C., Polanco X., Royauté J., Toussaint Y., INRIA research report N° 3198, juin 1997, available in postcript format, ftp.inria.fr (192.93.2.54)

Chapitre 7[1]

La conception de bases infométriques

Une application des programmes développés dans le cadre du projet HENOCH présenté dans le chapitre précédent est la possibilité de construire des bases de données infométriques hybrides (multi-sources, multi types de données) exploitables pour le calcul d'indicateurs de politique scientifiques selon un mode hypertexte. Rassemblant des informations scientifiques et techniques normalisées et codifiées, une base est dite 'infométrique' ou 'bibliométrique' lorsque sa structure a été conçue pour obtenir des indicateurs infométriques ou bibliométriques. Il n'existe pas de producteurs directs de bases infométriques mais des bases constituées à partir de données fournies par les producteurs de bases de données bibliographiques.

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

---

[1] Grivel L., Fagherazzi H. Fourneret P. Zerouki A. 'Conception de bases de données infométriques hybrides : analyse de la pratique de trois observatoires européens et propositions', Les systèmes d'information élaborée, Ile Rousse, Corse, Edition CD-ROM (CRRM - Marseille), 1999.

1 Introduction

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

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

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

L'objectif de l'article est double. Mettre en évidence quelques points clés et les difficultés pour construire ce type de base et tirer les leçons sur le plan informatique d'expériences[2] offrant une certaine similarité avec cette problématique. C'est pourquoi cet article comporte deux parties

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

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

2 Bases de données infométriques

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

2.1 Présentation des organismes et de leurs objectifs

a) L'Espagne

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

b) La France

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

c) La Hollande

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

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

2.2 Données et structure de données dans les bases infométriques

Nous mettrons l'accent dans cette sous-section sur ce qui caractérise une base infométrique. Les méthodes pour réaliser des indicateurs à partir de données bibliographiques vont de la statistique descriptive aux analyses multidimensionnelles, en passant par des techniques de classification et de cartographie ; [ROSTAING 96] constitue une bonne introduction à ces méthodes. L'aspect calcul et type d'indicateurs est abordé plus complètement dans [MOED 96], [GLANZEL 96].
On peut observer que la plupart des indicateurs publiés dans les rapports des trois observatoires étudiés sont des indicateurs univariés[3]. Les indicateurs relationnels les plus couramment utilisés sont les co-publications et cocitations, en se limitant à du dénombrement Les indicateurs les plus sophistiqués (classification, cartographie) ne sont employés que dans le cadre d'études à la demande (voir annexe II).

2.2.1 Données

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

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

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

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

Pourquoi les observatoires procèdent ils de cette manière ?

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

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

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

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

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

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

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

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

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

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

Les points les plus critiques sont selon les observatoires et dans cet ordre :

- une absence de politique claire concernant la couverture
- la saisie incomplète des auteurs,
- l'absence des citations.

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

2.2.2 Tables de nomenclatures / fichiers d'autorité

Rôle des fichier d'autorité : agréger et normaliser

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

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

Quelques exemples de fichiers d'autorité ou tables de nomenclatures

Disciplines/ domaines scientifiques

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

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

Entité géographique/institutionnelle

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

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

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

Espagne
Pour le traitement des affiliations, le CINDOC a constitué les fichiers d'autorité suivants :
I/-Centres de recherche
- Nom standardisé

- Code institution
  pour les centres espagnols à 5 niveaux :
  1. *dépendance administrative*

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

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

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

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

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

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

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

Facteur d'impact du périodique

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

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

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

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

2.3 Modélisation et stockage des données infométriques

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

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

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

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

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

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

2.4 Conclusion

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

Sur le plan méthodologique, les points clés sont :

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

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

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

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

Les autres obstacles sont de nature plus politique ou juridique. Par exemple, pour définir une couverture élargie, il est nécessaire d'interroger plusieurs bases de données. Certains

producteurs de données refusent ou font payer très cher la constitution de nouvelles bases à partir de données leur appartenant, imposant une licence à un coût élevé et/ou se donnant un droit de regard sur l'utilisation de ses données. Autre exemple : la constitution de fichiers d'autorités pour les organismes d'affiliation. Sans la collaboration des organismes concernés, il est difficile d'établir des fichiers pertinents. La fourniture d'un organigramme simplifie la tâche, de la même manière qu'il est plus facile de faire une normalisation des descripteurs (mots-clés) si on dispose de ressources terminologiques[4].

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

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

3 Intégration de données héterogènes

L'objectif de cette deuxième partie est de tirer les leçons de diverses expériences de veille[5] que nous avons menées. L'URI a développé une approche originale basée sur un couplage SGML/SGBD qui permet de construire et d'exploiter des indicateurs infométriques dans un environnement hypertexte convivial à des fins de veille scientifique, en employant une méthodologie un peu analogue à celle des observatoires des sciences et techniques (section 2) et des méthodes de traitement de données issues du monde de la gestion documentaire. Ces travaux ont débouché sur une plate-forme infométrique dont l'un des composants, le logiciel HENOCH, permet d'intégrer des données hétérogènes en types et en formats [GRIVEL 95,97,99], cf annexe 3).

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

Alimenter un SGBD à partir de documents fait partie des applications courantes dans le monde documentaire. D'une manière générale, il s'agit de transformer un document d'une certaine structure logique en une autre. L'intérêt de SGML/XML[6] dans ce contexte n'est plus à démontrer. On trouve aujourd'hui sur le marché plusieurs éditeurs SGML/XML disposant d'une interface avec les principaux SGBD du marché [MICHARD 98]. Il est ainsi possible, en utilisant les interfaces de programmation (API) de l'éditeur SGML/XML et du SGBD, de développer une passerelle de stockage dans la base de donnée de tout élément XML 'parsé' (analysé) par l'éditeur.

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

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

---

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

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

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

Nous exposons ici notre méthode et nous l'évaluons.

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

3.2.1 Reformatage

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

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

3.2.2 Intégration des données dans un SGBD : méthode

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

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

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

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

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

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

query :

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

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

Cette approche spécifie donc de manière déclarative les relations entre les éléments de données et leur représentation dans la base en utilisant une sorte de 'règle de réécriture' qui permet d'exécuter, par exemple une méthode de création d'un objet complexe (par exemple une super-notice[7] bibliographique) à partir des éléments de données.

3.3 Evaluation

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

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

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

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

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

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

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

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

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

4 Conclusion

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

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

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

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

BIBLIOGRAPHIE

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nb de réf citées

Laboratoires

20/5/1
12508319 Genuine Article#: LT747 Number of References: 52
Title: AMYOTROPHIC-LATERAL-SCLEROSIS AND STRUCTURAL DEFECTS IN CU,ZN SUPEROXIDE-DISMUTASE
Author(s): DENG HX; HENTATI A; TAINER JA; IQBAL Z; CAYABYAB A; HUNG WY; GETZOFF ED; HU P; HERZFELDT B; ROOS RP; WARNER C; DENG G; SORIANO E; SMYTH C; PARGE HE; AHMED A; ROSES AD; HALLEWELL RA; PERICAKVANCE MA; SIDDIQUE T
Corporate Source: NORTHWESTERN UNIV,SCH MED,DEPT NEUROL,300 E SUPER ST/CHICAGO//IL/60611; NORTHWESTERN UNIV,SCH MED,DEPT NEUROL,300 E SUPER ST/CHICAGO//IL/60611; SCRIPPS CLIN & RES INST,DEPT MOLEC BIOL/LA JOLLA//CA/92037; NORTHWESTERN UNIV,INST NEUROSCI/CHICAGO//IL/60611; UNIV CHICAGO,DEPT NEUROL/CHICAGO//IL/60637; DENT NEUROL INST,DEPT NEUROL/BUFFALO//NY/14209; DUKE UNIV,MED CTR,DEPT MED NEUROL/DURHAM//NC/27710; UNIV LONDON IMPERIAL COLL SCI TECHNOL & MED,DEPT BIOCHEM/LONDON SW7 2AZ//ENGLAND/; NORTHWESTERN UNIV,SCH MED,DEPT CELL MOLEC & STRUCT BIOL/CHICAGO//IL/60611
Journal: SCIENCE, 1993, V261, N5124 (AUG 20), P1047-1051

Source

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

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

Mots-clés obtenus par indexation automatique

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

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

.../...

Annexe 2 : Tableau comparatif des trois organismes étudiés

	OST Observatoire des Sciences et des Techniques 93, rue de Vaugirard 75006 PARIS Tél. : 01 42 22 30 30 Télécopie : 01 45 48 63 94	**CINDOC** Centro de Informacion y Documentacion Cientifica Joaquin Costa 22 28002 Madrid Tél : +34-1-5635482 Télécopie : +34-1-5642644	**CWTS** Centre for Science and Technology Studies Leiden University PO Box 9555 2300 RB Leiden Tel : +31 71 527 3909 Fax : +31 71 527 3911
Missions	« construire des indicateurs fiables, pertinents et pérennes, décrivant la science et la technologie françaises en comparaison européenne et internationale »	« élaboration de bases de données bibliographiques et réalisation d'analyses bibliométriques de la production scientifique espagnole, ainsi que normalisation de la terminologie scientifique »	cartographier la science et la technologie, plus particulièrement celles des Pays-Bas, en utilisant des méthodes quantitatives, spécialement des méthodes bibliométriques et infométriques.
Type d'organisme et effectif	Groupement d'Intérêt Public (GIP) de 14 membres : 7 ministères, 6 grands établissements publics (CEA, CNRS, CNES, CNET, INSERM, INRA) et l'ANRT. Membre associé : ORSTOM effectif environ 10 personnes	Centre de documentation scientifique du CSIC, (Consejo Superior de Investigaciones Cientifica). Environ 130 personnes	Centre financé par le NWO (Netherlands Organization for Scientific Research), 8 chercheurs, 4 ingénieurs, 2 secrétaires
Produits de l'organisme	***Publications*** : Indicateurs science et technologie, rapports annuels. La lettre de l'OST Les cahiers de l'OST ***Produits des ateliers de l'OST*** pour analyse stratégique à la demande (micro-indicateurs).	services comparables à ceux de l'INIST (fourniture de documents, recherches bibliographiques, traductions...), bases de données multidisciplinaires ICYT (science et technique) et ISOC (sciences humaines). Toutes ces bases de données couvrent spécifiquement la littérature espagnole. ***Concernant l'Infométrie*** - une base de données bibliométrique - une revue électronique *Cybermetrics* : journal international de recherche en scientométrie, bibliométrie et infométrie.	- une base de données bibliométrique. ***Publications*** - articles des chercheurs - rapports. ***Ex*** : rapport CWTS 98-01 (février 98) commandé par le ministère de l'éducation, de la culture et des sciences, sur la production et l'impact des Pays-bas dans les sciences humaines et sociales. - participe au rapport du NWOT publié tous les deux ans

Ressources	Pour calculer les indicateurs *bibliométriques* standards en sciences et techniques • les données du Science Citation Index (SCI), après extraction de certains journaux de psychologie et d'économie, enrichissement avec Compumath, produite elle aussi par l'ISI. • les bases EPAT et USPAT (brevets européens et américains enquêtes ministérielles, R.D. (recherche industrielle et innovation), MENDEP (étudiants et diplômés), OCDE, UNESCO, EUROSTAT (statistiques européennes), bases de données bibliographiques (PASCAL INSPEC, CHEMICAL ABSTRACT, SCI)	• Des bases de données bibliographiques (SCI, SSCI, ICYT, Physic Brief, INSPEC, Chemical Abstract, Biosis, MEDLINE, Exerpta Medica). • Des données factuelles : rapports officiels annuels et données de ressources humaines du monde scientifique et universitaire espagnol	Une base de données bibliométrique essentiellement constituée de publications scientifiques de chercheurs des Pays-bas dans les revues traitées pour SCI (Science Citation Index), SSCI (Social Science Citation Index), A&HCI (Arts & Humanities Citation Index) et publiées par l'ISI (Institute for Science Information). S'ajoutent à ces publications néerlandaises des données provenant des publications citant ces chercheurs pendant la même période.
Types d'indicateurs	*MACROINDICATEURS* : niveau d'observation à un niveau agrégé (pays, région), en comparaison internationale • mesure de niveau d'activité • indicateurs de spécialisation • indicateurs d'impacts • profils d'activité • copublications • cocitations • codépôt de brevet • matrices inventeurs-déposants de brevets *MICROINDICATEURS* : ciblés sur le plan géographique, institutionnel, produits à la demande	Macroindicateurs d'impact : Espagne en comparaison internationale • IF : Facteur d'impact moyen (pour une spécialité au niveau national) • RIF : Relative Impact Factor (comparaison internationale) Microindicateurs d'impact : comparaison des différents centres de recherches dans la même discipline Indicateurs de production scientifique par spécialité. Indicateurs de production scientifique par lieu. Copublications par spécialité. Copublications par lieu.	Sept types d'indicateurs : 1) Des indicateurs de production scientifique. 2) Des indicateurs d'impact. 3) Des indicateurs de positionnement sur les différentes revues scientifiques. 4) Des indicateurs d'orientation intellectuelle. 5) Des indicateurs de coopération ou de collaboration. 6) Des indicateurs de type de publication. 7) Des indicateurs de couverture en périodiques (revues scientifiques).

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

1 Description d' HENOCH

Le système HENOCH comprend:
1. un générateur de bases de données relationnelles à partir de documents au format SGML. Ce générateur utilise la notion d'arbre SGML comme structure pivot pour la description des données alimentant la base infométrique. Ces documents sont :
   a) les données initiales (qui sont de différents types et qui peuvent provenir de différentes sources : articles de périodiques, congrès, thèses, brevets) mises au format SGML et complétées (éventuellement) d'un certain nombre d'informations obtenues par traitements linguistiques (mot clés)
   b) les résultats de classification des données initiales (regroupement de documents ou de mots-clés) par les outils SDOC et NEURODOC [GRIVEL 95b],
   c) les tables de nomenclatures nécessaires pour la production de certains indicateurs.
2. un générateur des systèmes hypertextes sous WWW pour l'analyse, la valorisation et la diffusion des résultats de classification. Ce programme établit une interface WWW-SGBD par une passerelle qui permet de se connecter au SGBD, soumettre des requêtes SQL à partir d'un modèle de page HTML incluant des requêtes SQL, récupérer le résultat et le mettre au format HTML conformément au modèle, et enfin se déconnecter.

Le générateur de base relationnelle procède en deux étapes :

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

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

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

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

2 La fusion de données multi-sources

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

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

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

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

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

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

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

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

Chapitre 8[1]

L'analyse de l'IST sous HENOCH : une illustration dans le domaine des plantes transgéniques

Le processus d'analyse de l'information est un mélange d'exploration informelle intuitive et d'exploitation méthodique de l'information élaborée par différents outils d'analyse. Ce chapitre montre par un jeu de questions-réponses comment un hypertexte généré par le système présenté au chapitre 6 permet à ses usagers, par exemple un chercheur, de découvrir les thématiques à la frontière de son domaine de recherche, les équipes qui travaillent sur le même sujet que lui, des revues dans lesquelles publier, des congrès dans lesquels publier et auxquels assister.

L'information est organisée sous la forme d'un hypertexte basée sur une métaphore cartographique. Ainsi l'utilisateur dispose d'outils de navigation qui permettent d'éviter le phénomène de désorientation commun aux hypertextes. Pour naviguer, l'utilisateur dispose d'une carte, d'une "boussole" pour orienter sa carte (sa connaissance du domaine) et de méthodes pour faire le point, connaître son positionnement et celui des autres.

Deux types de navigation complémentaires sont proposés :

- une exploration intuitive basée sur la carte thématique permettant d'accéder rapidement à des listes pondérées de mots-clés, auteurs, affiliations, sources pour chaque thème, puis de naviguer vers les documents associés à chaque élément de ces listes.

- des fonctions de recherche basées sur ces indicateurs permettent par exemple de savoir dans quelles thèmes un organisme est positionné, le nombre de documents qui est à l'origine de ce positionnement dans le corpus pour chaque thème, puis de naviguer vers ces documents.

L'usager dispose donc de plusieurs modes de navigation conviviaux lui permettant de satisfaire ses multiples besoins:

- avoir une vue d'ensemble,

- suivre et analyser l'évolution thématique, identifier des relations inter-thèmes non explicites,

- repérer l'émergence de nouveaux thèmes de recherche ,

- identifier et regrouper les acteurs, les institutions, leurs vecteurs de communication (thèses, rapports, monographies, périodiques) par thèmes

- évaluer le positionnement thématique d'un acteur, d'une institution, d'un pays, d'un vecteur de communication (périodique, congrès, ...).

Ces besoins sont illustrés dans le cadre d'une étude sur les plantes transgéniques.

---

[1] Grivel L. 'HENOCH, un outil d'analyse de corpus d'information scientifique et technique', Le Micro Bulletin Thématique n°3, L'information scientifique et technique et l'outil Internet, CNRS-DSI, p.27-44, 1999.

1 Présentation générale d'HENOCH

Henoch est le résultat de travaux menés au sein de l'Unité Recherche et Innovation de l'INIST. Cet article définit le public visé par cet outil, son organisation des données selon des principes infométriques pour répondre aux besoins énoncés, puis tente de montrer, par un jeu de questions-réponses, comment se système peut aider des organisations (laboratoire, entreprises) dans leur travail d'exploration et d'analyse de l'information scientifique relative à leur domaine d'activité.

1.1 A qui s'adresse HENOCH ?

HENOCH s'adresse aux chercheurs, veilleurs, spécialistes d'un domaine scientifique, technique ou économique non nécessairement professionnels de la documentation ou de l'informatique, documentalistes qui**, sur un sujet ou un domaine donné** , veulent, à partir des bases bibliographiques PASCAL et FRANCIS[2] :

- avoir une vue d'ensemble,
- suivre et analyser l'évolution thématique, identifier des relations inter-thèmes non explicites,
- repérer l'émergence de nouveaux thèmes de recherche ,
- identifier et regrouper les acteurs, les institutions, leurs vecteurs de communication (thèses, rapports, monographies, périodiques) par thèmes
- évaluer le positionnement thématique d'un acteur, d'une institution, d'un pays, d'un vecteur de communication (périodique, congrès, ...).

Autrement dit explorer et analyser l'information relative à leur sujet de préoccupation (un corpus bibliographique) pour, par exemple:

- avoir une première approche d'un sujet de recherche,
- orienter des recherches,
- identifier des technologies émergentes,
- évaluer les résultats d'une équipe de recherche,
- établir un partenariat, ...

Habituellement un corpus bibliographique sur un sujet ou un domaine donné peut représenter quelques milliers de références qu'il est exclu de parcourir séquentiellement. Dans HENOCH, un tel corpus est structuré selon des principes infométriques de manière à constituer une bases de données dites infométriques, exploitables pour l'analyse de l'information.

1.2 Qu'est ce qu'une base de données infométriques, à quoi ça sert ?

Pour permettre cette analyse de l'information, HENOCH exploite des indicateurs. Ces indicateurs sont le résultat d'un ensemble de traitements linguistiques et statistiques (classification et cartographie) appliqués à des données structurées de type références bibliographiques ou brevets représentatifs d'un domaine

---

[2] HENOCH peut fonctionner à partir de données provenant d'autres bases, mais seules des données provenant de nos bases seront accessibles par Internet.

Ce sont :

1. les mots-clés comme indicateurs de la connaissance véhiculée par le document, associés aux références
bibliographiques de façon manuelle ou assistée par ordinateur ;

2. les classes comme indicateurs des thèmes ou centres d'intérêt autour desquels s'agrège l'information (articles, auteurs, institutions, périodiques) ;

3. et enfin, la carte comme indicateur stratégique de la position relative des thèmes dans l'espace de connaissance
couvert par les documents analysés.

En conclusion, nous définissons une base de données infométriques comme rassemblant et structurant ces informations en les associant aux données bibliographiques, donnant ainsi la possibilité de mettre en relation tout élément constitutif d'une référence bibliographique (auteur, titre, source, affiliation,...) avec les thèmes (classes ou clusters) obtenus par classification automatique.

Les caractéristiques concernant la classification et la cartographie sont disponibles à l'adresse de l'*Unité Recherche et Innovation* "http://www.inist.fr/pri/pri.htm" et dans un article publié dans la revue électronique SOLARIS accessible également sur Internet (http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2grivel.html).

Un compte rendu d'un séminaire tenu à l'ADEST le 9.12.97, http://www.upmf-grenoble.fr/adest/seminaires/francois.htm fait le point sur les éléments techniques composant la plate-forme infométrique.

1.3 Architecture informatique

D'un point de vue informatique, HENOCH réalise une passerelle entre trois éléments:

- un système infométrique (extracteur terminologique, moteurs de classification et de cartographie),
- un système de gestion de bases de données relationnel (ici Oracle),
- un navigateur sur le Web,

en assurant deux fonctions principales :

- alimenter le SGBD à partir de documents SGML produits par le système infométrique,
- établir une interface WWW-SGBD par une passerelle qui sait donc se connecter au SGBD, soumettre des requêtes SQL à partir d'un modèle de page HTML incluant des requêtes SQL, récupérer le résultat et le mettre au format HTML conformément au modèle, et enfin se déconnecter.

Le cadre informatique de ce développement a été plus précisément décrit dans [1,2,3].

2 Comment HENOCH permet d'explorer et d'analyser l'information scientifique et technique sans avoir à faire l'apprentissage de commandes complexes ?

L'information est organisée sous la forme d'un hypertexte basée sur une métaphore cartographique. Ainsi HENOCH dispose d'outils de navigation qui permettent d'éviter le phénomène de désorientation commun aux hypertextes. Pour naviguer, l'utilisateur dispose d'une carte, d'une "boussole" pour orienter sa carte (sa connaissance du domaine) et de méthodes pour faire le point, connaître son positionnement et celui des autres. C'est le rôle joué par les indicateurs infométriques.

HENOCH propose deux types de navigation complémentaires en exploitant les indicateurs infométriques :

- une exploration intuitive basée sur la carte thématique permettant d'accéder rapidement à des listes pondérées de mots-clés, auteurs, affiliations, sources pour chaque thème, puis de naviguer vers les documents associés à chaque élément de ces listes.

- des fonctions de recherche basées sur ces indicateurs permettent par exemple de savoir dans quelles thèmes un organisme est positionné, le nombre de documents qui est à l'origine de ce positionnement dans le corpus pour chaque thème, puis de naviguer vers ces documents.

L'interface d'HENOCH obéit au principe des interfaces métaphoriques, c'est à dire qu'elle permet à l'utilisateur de travailler sans nécessiter l'apprentissage fastidieux de procédures et de commandes.

Nous prendrons comme exemple un corpus issu de la base Pascal (1 339 enregistrements (production Pascal 1995-96) qui a été utilisé par Le Bureau Van Dijk et l'INIST afin de réaliser un rapport de tendance sur les plantes transgéniques.

L'outil de classification et cartographie qui a été employé est NEURODOC. HENOCH, en organisant les résultats du programme précédent dans un SGBD, a permis aux équipes BVD et INIST de réaliser l'analyse de l'information collectée à partir de son interface WWW dont nous illustrons ici les fonctionnalités.

Après s'être connecté à HENOCH et avoir choisi son corpus de travail, l'utilisateur se trouve face à une page contenant une barre de menu principal et la liste de thèmes qui ont été constitués automatiquement.

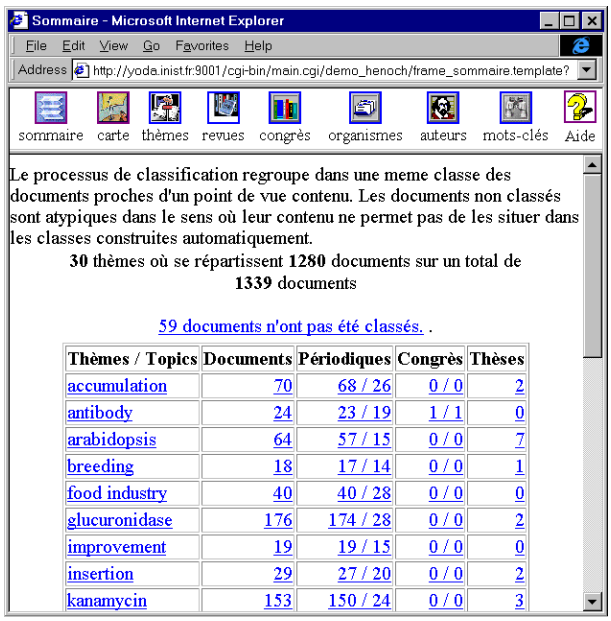


Figure 1

La barre de menu principale est composée de 8 boutons plus une aide en ligne. Les trois premiers boutons permettent d'accéder respectivement à la liste des bases de données infométriques disponibles, à la carte thématique, au tableau des thèmes (Fig. 1), les 5 suivants correspondent aux fonctions de positionnement. Nous allons voir leur utilisation, en commençant par la carte thématique qui est un peu le poste de pilotage de la navigation tandis que le tableau des thèmes présenté en premier lieu a surtout pour objectif de fournir à l'utilisateur des informations quantitatives sur les résultats de classification.
Le lecteur peut notamment se faire rapidement une idée de la distribution des documents dans les thèmes et par type de document. Par exemple, le thème « Accumulation » rassemble 70 documents dont 68 répartis dans 26 périodiques et 2 thèses.

2.1 Comment naviguer depuis la carte thématique ?

Depuis la carte de thématique (Fig.2), l'analyste peut accéder à différents types d'informations pertinentes pour un thème et visualiser très rapidement les éléments les plus représentatifs du thème, son organisation.

Pour l'utilisateur, la procédure est la suivante:

1) Choisir, dans la table de boutons radio à gauche de la carte, un type destination

MC --> une liste triée de mots-clés,
TI --> une liste triée de titres des documents associés,
AU --> une liste triée des auteurs,
AF --> une liste triée des affiliations des auteurs,
SO --> une liste triée des des modes de publication des auteurs (les sources),

2) Choisir un thème sur la carte (en cliquant sur le nom du thème ou sur le cercle noir positionné avant le nom.)

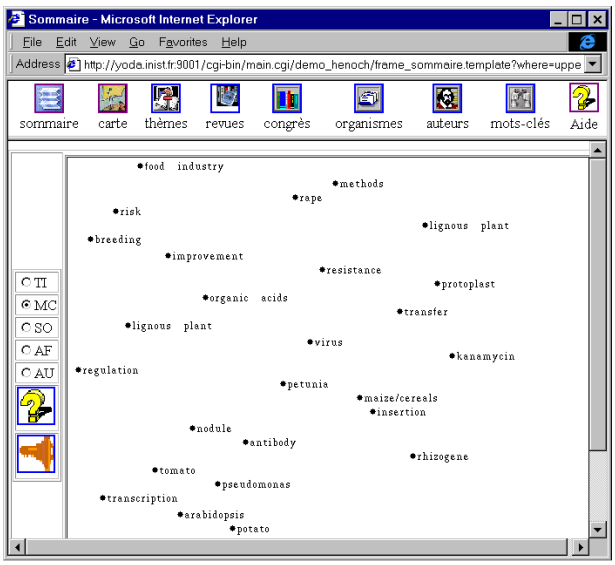


Figure 2

2.2 Comment analyser la carte ?

L'analyse de la carte dépend de la méthode de cartographie employée. Dans le cas présent, l'algorithme de cartographie, l'analyse en composante principale, réduit le nombre de dimensions de données statistiques multidimensionnelles, de telle manière que deux thèmes au contenu similaire soient relativement proches sur la carte. L'interprétation de la carte est délicate car le lecteur doit s'appuyer sur les connaissances qu'il a du domaine et sur une observation attentive de chaque thème (opération illustrée dans 4.3) L'interprétation de la carte correspondant à ce jeu de données est disponible dans le rapport de tendance qui est commercialisé.

2.3 Comment observer l'organisation thématique ?

Observer l'organisation d'un thème suppose de pouvoir décrire son contenu, les auteurs qui travaillent sur ce thème, leur organisme d'appartenance, leur modes de publication, les relations avec les autres thèmes.

L'organisation du thème est décrite par :

- une liste de mots-clés ordonnés selon leur importance pour la définition du thème, le mot-clé de plus fort poids donnant son nom au thème
- une liste de titres de documents ordonnés selon le même critère,
- une liste des affiliations, ensemble des affiliations des auteurs des documents du thème, ces derniers sont triés selon leur fréquence dans le thème,
- une liste d'auteurs : ensemble des auteurs des documents du thème, ces derniers sont triés selon leur fréquence dans le thème et
- une liste de sources : ensemble des titres de revues où sont édités les documents du thème, ces dernières sont triés selon leur fréquence dans le thème.

On peut accéder à chacune de ces informations par la carte. Il existe une fenêtre par type d'information (mots-clés, titres des documents, auteurs, affiliations, sources) associé à chaque thème. Une barre de menu commune à toutes ces fenêtres et locale au thème permet également d'accéder à ces différents types d'informations sans repasser par la carte et informe sur la quantité d'information qui est agrégée autour du thème. Les

documents du thème pouvant appartenir à d'autres thèmes, un lien "Documents partagés avec d'autres thèmes" permet d'accéder à la distributions des documents par thème.

2.3.1 Comment se faire une idée du contenu d'un thème ?

Deux moyens différents sont mis à disposition de l'utilisateur via le menu local au thème:

- la liste triée de mots-clés,
- la liste triée des documents du thème.

2.3.1.1La liste triée de mots-clés (Fig.3)

Le lien **Description** renvoie donc à la liste triée de mots-clés ordonnés selon leur importance pour le thème. Chaque mot-clé est précédé de son poids et de sa fréquence locale et globale. Le mot de poids le plus élevé donne par défaut son nom au thème.

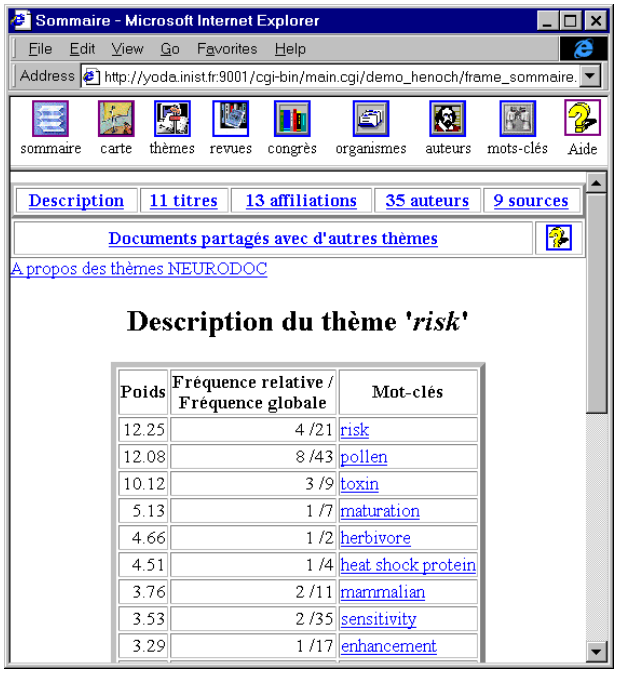


Figure 3

La colonne Fréquence globale donne le nombre total de documents indexés par chaque mot du thème dans le corpus. Il ne s'agit pas de la fréquence du mot dans les documents relatifs au thème, fréquence dite locale. Leur rapport donne le pourcentage relatif de documents indexé par le mot dans le thème.

L'utilisateur peut donc visualiser rapidement les mots-clés liés à ce thème triés par degré de pertinence par rapport au thème et utiliser chacun des mots pour effectuer des recherches locales au thème « risk ». En effet chaque mot-clé donne accès à la liste des titres des documents du thème qui sont indexés par ce mot dans le thème (Fig.4). Ce qui peut lui donner des idées sur des termes liés à « risk » et l'aider à formuler de nouvelles hypothèses. Exemple : «Pollen » et « heat shock protein ».
Sur 21 documents indexés par « risk », 4 seulement figurent dans le thème, alors que le thème ne comporte en tout que 11 documents correspondant globalement à ce profil thématique.

Ainsi donc le mot « risk » a été employé dans différents contextes. Pour les 4 documents en question, il s'agit plutôt des risques pour l'environnement à travers les pollens.

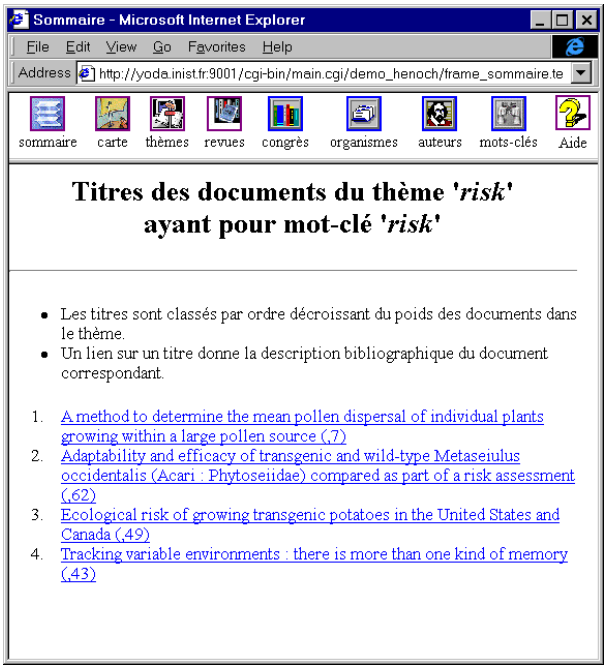


Figure 4

2.3.1.2 La liste triée des titres de document (Fig. 5)

Le lien **titres** renvoie à la liste de titres documents qui est ordonnée selon la pertinence (le poids) du document pour le thème. La date de publication et le poids de chaque document encadrent le titre en donnant accès à la description bibliographique du document correspondant.
Ces deux éléments (date et poids) permettent au lecteur de se faire une idée de l'âge moyen d'un thème, puisque les documents les plus pertinents (d'un point de vue statistique) pour le thème sont classés en tête. A partir de ces indicateurs, l'utilisateur peut faire des hypothèses sur les thèmes en vogue (nombre important de documents, date récente pour l'essentiel des documents), en perte de vitesse (faible nombre de documents, date ancienne), en émergence (faible nombre de documents, date récente). Bien entendu, c'est l'expertise du lecteur, ses recoupements avec d'autres informations qui lui permettront d'affirmer ou d'infirmer la validité de ses hypothèses.

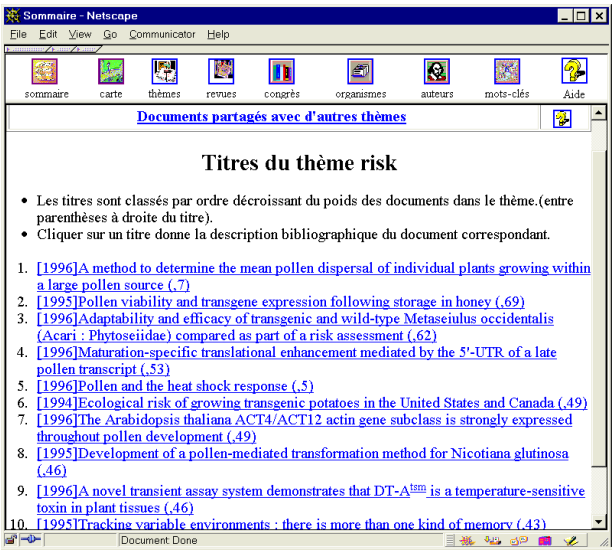


Figure 5

2.3.2 Comment se faire une idée de l'environnement institutionnel d'un thème ?

Le lien **affiliations** renvoie à la liste des affiliations des auteurs des documents du thème (Fig. 6), triées selon leur fréquence dans le thème. Ceci donne une indication de la productivité de l'institution dans le thème.
Chaque affiliation est précédée de sa fréquence dans le thème et donne accès à la liste des titres des documents du thème écrits par des auteurs membres de cette affiliation.

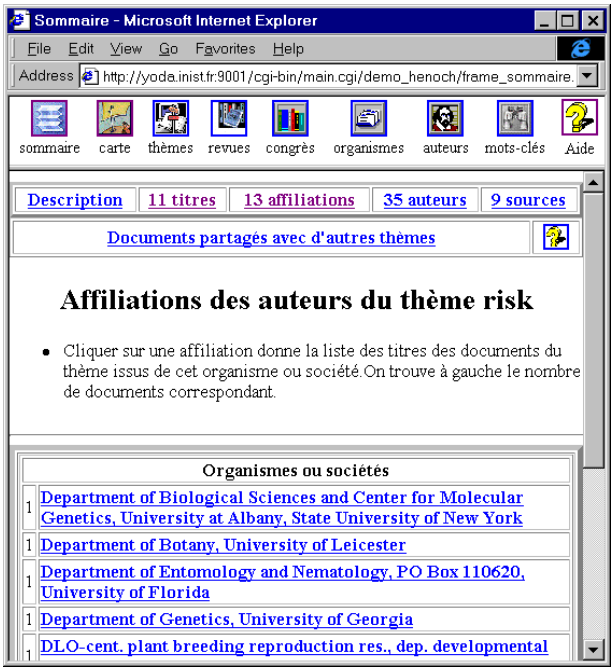


Figure 6

2.3.3 Comment se faire une idée des acteurs principaux du thème ?

Le lien **Auteur** renvoie à la liste des auteurs des documents du thème, triés selon leur fréquence dans le thème (Fig. 7).

Chaque auteur est précédé de sa fréquence dans le thème (sa productivité dans le thème) et donne accès à la liste des titres des documents du thème écrits par l'auteur.

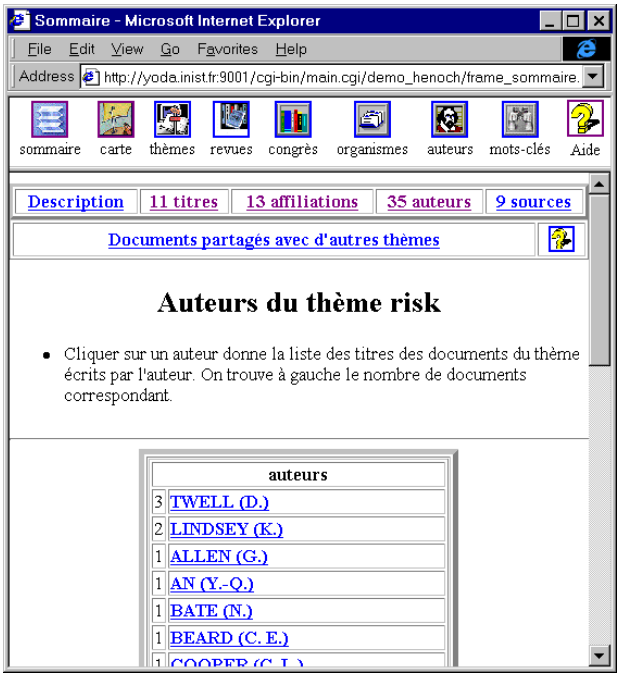


Figure 7

2.3.4 Comment se faire une idées des modes de publications des auteurs du thème ?

Le Lien **source** renvoie aux listes des titres de revues, congrès ou universités de soutenance pour les thèses, triées selon leur fréquence dans le thème, où sont publiés les documents du thème (Fig. 8). Les sources sont précédées de leur fréquence et donnent accès la liste des titres des documents du thème publiés par cette source. La distribution selon le type de source(revue, congrès) , puis pour chaque type de source selon le journal ou le nom du congrès permet de qualifier les modes de communications privilégiés des auteurs.
Exemple

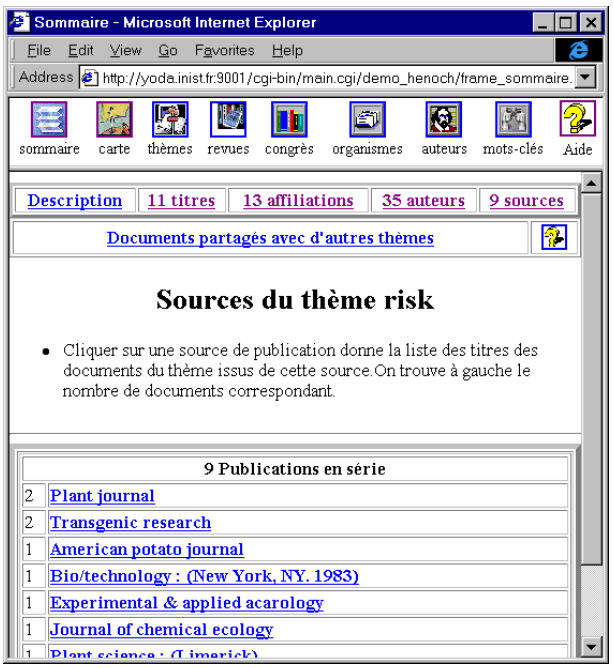


Figure 8

2.3.5 Comment se faire une idée des relations qu'un thème entretient avec les autres thèmes ?

Le lien **Documents partagés** permet d'accéder en premier lieu à la distribution des documents dans les autres thèmes, puis pour chaque thème, à la liste des titres des documents communs (Fig. 9). Le titre renvoie à la description bibliographique du document correspondant.

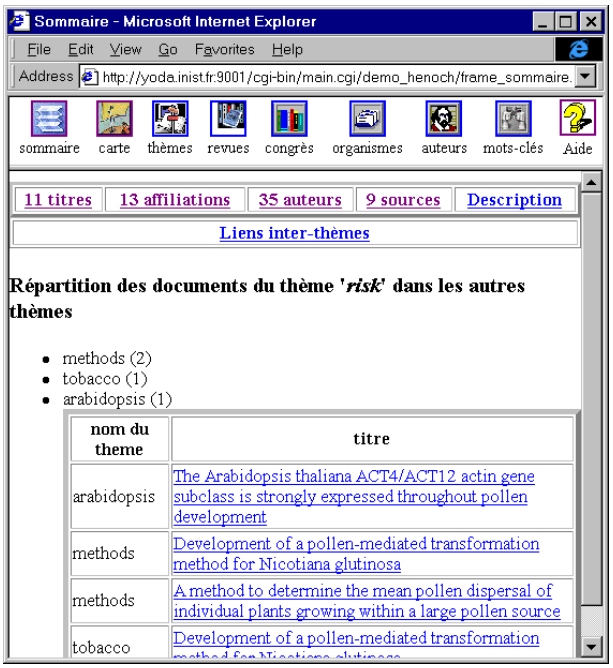


Figure 9

2.4 Comment utiliser la description bibliographique d'un document ?

La figure 10 montre un exemple de référence sélectionné à partir de la liste des documents d'un thème. Les flèches de navigation donnent la possibilité de naviguer transversalement vers des documents voisins d'un poids immédiatement supérieur ou inférieur dans le thème et la possibilité d'accéder aux mots clés décrivant le thème ainsi que de revenir à la liste des documents du thème. Un document peut en effet se trouver dans plusieurs thèmes, comme dans le cas ci-dessous.

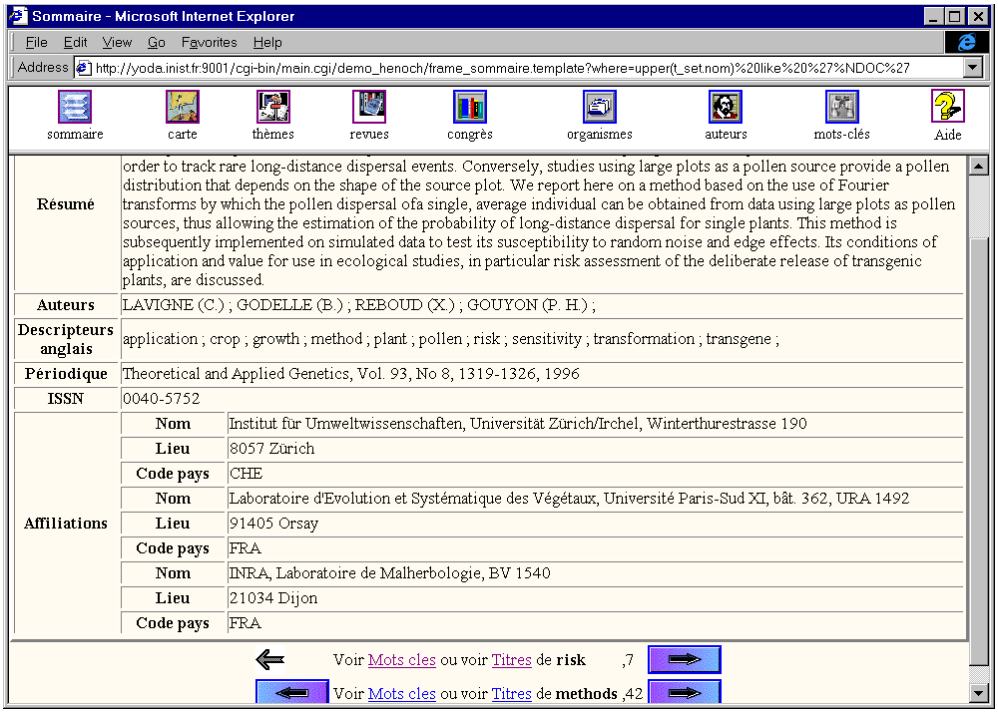


Figure 10

L'accès à la description bibliographique du document permet de compléter les observations effectuées et peut suggérer de nouvelles voies de navigation. Ici, supposons que l'utilisateur s'interroge sur le positionnement thématique de la revue « Theoretical and Applied Genetics » puis sur le positionnement des 3 organismes qui coopèrent et enfin se faire une idée des contextes dans lequel le mot-clé «risk » est employé. La section suivante illustre comment ce besoin peut être satisfait.

2.5 Comment effectuer le positionnement d'un périodique (d'un auteur, d'une affiliation, d'un mot-clé) dans les thèmes ?

Objectif

Le but est de savoir dans quelles thèmes un périodique (un congrès, un auteur, une affiliation, un mot-clé) est positionné, le nombre de documents qui est à l'origine de ce positionnement dans le corpus pour chaque thème, puis de naviguer vers ces documents.

Procédure à suivre

Il faut d'abord savoir si le périodique (le congrès, l'auteur, l'affiliation, le mot-clé) dont on cherche le positionnement thématique est bien dans la liste des périodiques (des congrès, des auteurs, des affiliations, des mots-clés) et sous quelle(les) forme(s) il a été saisi.

Chacun des boutons 4 à 8 qui figurent de manière permanente dans le menu principal, a pour effet de d'afficher une boite de sélection composée d'une zone de saisie et d'un bouton intitulé "filtrer" qui permet de faire une recherche sur la liste correspondant à l'intitulé du bouton (revues, congrès, organismes, auteurs, mots-clés).

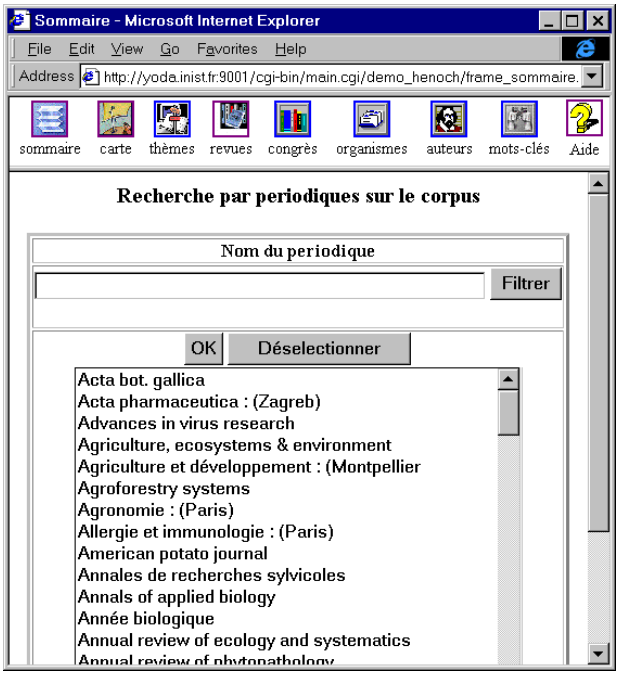


Figure 11

Quand le nombre d'éléments est important, les lister comme dans la figure 11 peut prendre du temps. C'est pourquoi, dans le cas d'une liste de plus de 1000 éléments, HENOCH n'affiche pas directement tous les éléments de la liste. A la place, est proposé la possibilité de filtrer par une expression entrée dans une zone de saisie. Par exemple en entrant les premières lettres de l'élément en utilisant la troncature à droite (le caractère *). Les minuscules et les majuscules ne sont pas différenciés.

Par exemple, *Genetics* signifie tout élément contenant "genetics".
a* donnera tous les éléments de la liste commençant par a.
Si le lecteur veut malgré tout afficher toute la liste, il doit taper * dans la zone de saisie puis cliquer sur le bouton "filtrer".

Ensuite, il faut sélectionner le ou les éléments intéressants dans la liste proposée (titre de périodique, congrès, auteur, affiliation, mot-clé) et valider en cliquant sur le bouton OK.

Le résultat est une distribution des documents relatifs à ce ou ces éléments par thème.
En sélectionnant un nom de thème et en validant par OK, on accède à la liste des titres des documents relatifs à ce thème pour le ou les éléments sélectionnés.

Dans le cas de la revue qui nous intéresse (« Theoretical and Applied Genetics »), le lecteur peut voir (Fig. 12) qu'elle se positionne en premier lieu dans le thème « Protoplasts » qui correspond au transfert de gènes dans les protoplastes et régénération des plantes à partir de cultures de protoplastes.

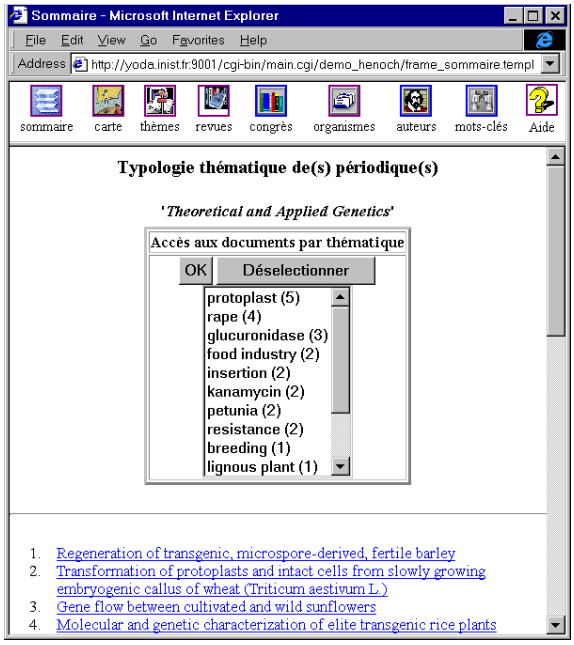


Figure 12

Pour des raisons diverses, les titres de périodique, les noms de congrès ou d'auteurs ou les affiliations peuvent se trouver sous différentes formes lexicographiques. En effet, il n'est pas possible de normaliser de manière fiable ces entités sans utiliser de fichiers d'autorité. Si on prend l'exemple du Laboratoire de Malherbologie, on peut observer qu'il figure dans la base de données sous 3 formes différentes (Fig.13). Ce qui ne pose pas de problème pour HENOCH, car on peut sélectionner plusieurs formes lexicographiques en les considérant comme constituant un objet unique.

Figure 13

Le lecteur peut donc observer le positionnement thématique de ce laboratoire (Fig.14) et lister les documents qu'il a produit. Dans le cas présent, ce laboratoire est à l'origine de trop de documents dans le corpus pour en tirer des conclusions.

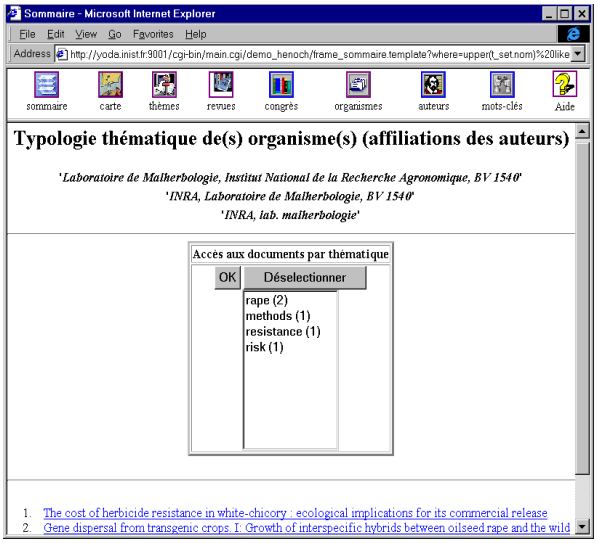


Figure 14

Le même type d'opération peut être effectué pour un mot-clé (Fig.15).

Figure 15

Les documents indexés par le mot-clé « risk » se répartissent entre différents thèmes : « risk » (les risques encourus par les consommateurs et l'environnement), « food industry » (l'agro-alimentaire) ,...
Le lecteur peut ensuite lister les documents thèmes par thèmes.

3 Conclusion et perspectives

Dans sa version prochainement accessible sur Internet, HENOCH, offrira l'accès à un ensemble de corpus bibliographiques constitués à partir des bases PASCAL OU

FRANCIS sur quelques sujets sélectionnés par l'INIST comme représentatifs de préoccupations de recherche actuelles.

De l'observation des réactions des utilisateurs dépendront les orientations que prendra l'INIST concernant l'exploitation d'un système tel qu'HENOCH.

Par ailleurs, depuis plusieurs années, l'Unité Recherche et Innovation de l'INIST est sollicité par divers organismes publics et privés qui souhaitent que leurs données (scientifiques et techniques pour la plupart) soient traitées par des outils linguistiques et statistiques. Ils souhaitent également que ces données et ces résultats soient rendus accessibles, via le réseau, à la communauté de leurs membres à travers une interface conviviale comme celle d'HENOCH.

Parallèlement, on voit se mettre en place, ici ou là, la notion d'observatoire thématique produisant de l'information élaborée partagée par un cercle d'utilisateurs-souscripteurs, les thèmes répondant aux besoins exprimés par un groupe d'utilisateurs, (les membres d'un organisme, une association, ...), ou à l'initiative d'un producteur d'information sur un thème porteur.

On peut donc imaginer dans le futur un lieu virtuel (village Internet ?) où l'information partagée autour d'une thématique serait exclusivement réservée aux membres du cercle, lesquels pourraient également commander les documents en ligne, demander une études complémentaire ou échanger des informations via un forum spécialement mis à leur disposition. Un système comme HENOCH pourrait donc tout à fait s'intégrer dans un bouquet de services.

4 BIBLIOGRAPHIE RELATIVE à HENOCH

1. Grivel L., X. Polanco, A. Kaplan "A computer System for Big Scientometrics at the Age of the World Wide Web", Scientometrics, vol.40, N°3, 1997, 493-506, 1997, et également in proceedings of the 6th International Conference on Scientometrics and Informetrics, Jerusalem, 131-142,1997.

2. Grivel L., C. Francois, X. Polanco -" Analyse de l'information par cartographie neuromimétique et requêtes SQL sur le Web", - "4ème Conf. Intern. Hypertextes et Hypermedias : réalisation, outils méthodes", *Hypertextes et Hypermedias,* Editions Hermès, Vol.1, n°2 ,237-248, 1997.

3. Grivel L., X. Polanco, A. Kaplan -"Requêtes et navigation à partir de l'information structurée, le système HENOCH ", Le Micro Bulletin, N°70, 493-506, 1997.

Chapitre 9

Bilan critique et perspectives

Après dix ans de travail de recherche, il est nécessaire de jeter un regard critique sur ce que l'on a réalisé. Ce dernier chapitre permet, à partir d'une évaluation critique des fonctions du système par un groupe d'utilisateurs, de dégager diverses voies de recherches possibles, notamment la visualisation et la comparaison dans le temps de représentations cognitives de données, la classification incrémentale, qui constituent de nouveaux enjeux pour la recherche sur la génération automatique d'hypertextes ergonomiques.

Tout au long des articles constituant le corps de cette thèse, il a été souligné l'importance d'une exploitation coordonnée de différentes techniques pour analyser l'Information Scientifique et Technique (IST), telle qu'elle est représentée dans les grandes bases bibliographiques. Deux outils (SDOC et NEURODOC) qui permettent de classer et représenter sur une carte un ensemble de documents en se basant sur les mots-clés, descripteurs du contenu des documents, ont été mis en œuvre et étudiés en profondeur, sur différents domaines d'application (chapitres 2, 3, 4 et 5). Ces études ont montré que l'exploitation et l'interprétation des résultats obtenus par de tels outils d'analyse supposent un mélange d'exploration informelle intuitive et d'exploitation méthodique de l'information élaborée par ces outils d'analyse. En partant d'une métaphore, la navigation dans un océan d'informations, il a été établi la nécessité de générer automatiquement des hypertextes, avec leur carte de navigation et des indicateurs de positionnement, à partir des données à analyser. L'exploration de cette voie a débouché sur la conception et le développement d'un système informatique, HENOCH qui permet de rassembler et d'organiser dans un SGBD (Système de gestion de bases de données) des données bibliographiques normalisées et traitées par diverses techniques, puis de distribuer ces informations sur INTERNET via une interface de navigation générée automatiquement et adaptée à l'analyse de l'information (chapitres 6, 7 et 8).

Sur le plan technologique, l'originalité d'HENOCH est de s'appuyer sur SGML pour réaliser le couplage SGBD/Web. Ce couplage permet non seulement l'intégration de données hétérogènes (des notices dans différents formats, des résultats de classification, des tables de nomenclatures, etc.) dans une base, mais aussi de distribuer des informations extraites de la base de données sous forme de données SGML ou HTML, soit pour des traitements ultérieurs, soit pour naviguer dans la base infométrique à travers une interface hypertexte dont les liens sont exprimés dynamiquement sous forme de requêtes dans le SGBD.

Sur le plan conceptuel, il a été montré expérimentalement que l'hypertexte, en tant que principe d'organisation de l'information, permet de modéliser et de mettre en place concrètement des mécanismes d'exploration et les interactions nécessaires entre les schémas mentaux de l'utilisateur (sa représentation du domaine couvert par la littérature scientifique) et différentes représentations fournies par les méthodes d'analyse employées.

Enfin et surtout, il y a le retour positif des utilisateurs concernant l'utilisation des hypertextes générés par HENOCH, notamment sur le plan de l'adaptabilité et l'ergonomie (section 1). Mais quelques points doivent être mieux pris en compte pour que cet outil réponde pleinement aux besoins de veille scientifique (section 2). L'analyse de ces manques ou faiblesses permet d'envisager quelques pistes d'améliorations qui constituent autant de perspectives de recherche (section 3).

1 Les points forts : adaptabilité et ergonomie

Un groupe de personnes à l'INIST de différents profils (informaticien spécialiste du Web, ingénieurs documentalistes, veilleur concurrentiel) a effectué une évaluation des fonctionnalités proposées et de l'interface. Cette dernière a été jugée agréable d'utilisation et facile à appréhender. Elle permet de travailler sans nécessiter l'apprentissage fastidieux de procédures et de commandes. Sur le plan des fonctionnalités, une évaluation sur le fond (leurs besoins versus les fonctions réalisables) met en évidence une liste de points forts/points faibles.

En résumé, les points forts sont :

1. une vue d'ensemble du corpus et de son organisation thématique (niveau corpus),
2. une vision d'un domaine par le biais des mots-clés,
3. la connaissance des acteurs liés à un thème,
4. la recherche des sources pertinentes,
5. la possibilité de savoir où publient les auteurs significatifs,
6. une évaluation quantitative des forces engagées derrière chaque thème (nombre d'auteurs, nombre d'organismes).

Pour un ingénieur documentaliste, cela signifie la possibilité de définir un vocabulaire pertinent pouvant améliorer les vocabulaires d'indexation ou de recherche, ou l'aider à la construction d'un plan de classement, un thésaurus.

Pour un chercheur, ces points forts se traduisent en la possibilité de découvrir les thématiques à la frontière de son domaine de recherche, le nombre et le nom des équipes qui travaillent sur le même sujet que lui, des revues dans lesquelles publier, des congrès dans lesquels publier et auxquels assister, etc.

Si on projette ces fonctionnalités à l'échelle d'un laboratoire ou d'un département scientifique, de telles bases de données structurées par thèmes constituent une mine d'informations partagées par les membres du laboratoire pour effectuer une réflexion stratégique sur les axes de recherche du laboratoire (ses forces et faiblesses, le positionnement de ses concurrents).

2 Les points faibles : la détection et l'analyse des évolutions thématiques dans le temps

Les apports d'un tel environnement sont jugés faibles ou insuffisants concernant les objectifs suivants :

- repérage de nouvelles orientations, voire de nouvelles activités ou de nouvelles collaborations d'un acteur traditionnel,
- suivi de l'évolution d'un thème,
- repérage de nouveaux acteurs, services ou produits,
- identification des tendances par rapport à un marché.

Cette liste de points faibles constitue en fait une même problématique. A ce jour, HENOCH permet de visualiser des photographies (classifications) successives des corpus de données sans qu'il y ait des moyens objectifs de mesurer l'évolution entre deux photographies. Comment dépasser ces limites et détecter des évolutions au fil du temps, détecter des signaux faibles, des tendances ?

Considérant la difficulté à comparer des cartes thématiques dont les thèmes ont changé E. Noyons et A. Van Raan ont proposé récemment deux type de comparaison dans le temps des cartes thématiques [Noyons 1998] :

- reconstruire le présent à partir du passé: le principe est d'affecter les articles publiés durant l'année t à une classification d'articles d'une année antérieure (par exemple t-4), puis d'observer l'évolution des proximités thématiques sur deux cartes (au temps t et t-4 par exemple) obtenues par la méthode du Multi Dimensional Scaling (MDS)

[KRUSKAL 1964] qui est une méthode de cartographie planaire qui tente de respecter aux mieux les distances entre points voisins dans un espace multi-dimensionnel.

- Re-visiter le passé à partir du présent : c'est à dire affecter les articles publiés durant les années antérieures à une classification d'articles de l'année présente, puis reconstruire une carte par la méthode du MDS puis d'observer l'évolution entre les deux cartes. C'est en fait ce mode de comparaison qui semble le meilleur, car, bien évidemment la situation présente est mieux décrite.

L'inconvénient de cette méthode est aussi son principal avantage. Il est plus facile à effectuer la comparaison entre deux cartes car les noms des thèmes sur la carte n'ont pas changés (il s'agit de la même classification), mais bien évidemment, on ne peut pas observer les évolutions entre classifications.

Une autre approche s'inspirant des méthodes utilisées en intelligence artificielle pour la représentation des connaissances pourrait être utilisée [GODIN et al. 1998].

Dans ce type de méthode, la hiérarchie des classes est restructurée (calculée incrémentalement) grâce à des opérateurs chaque fois qu'une nouvelle information est soumise au système. Ainsi, à chaque étape, la nouvelle donnée est comparée avec les classes déjà construites. L'arrivée d'un nouvel élément peut aussi bien avoir un effet modéré (faire grossir une classe existante) que provoquer un bouleversement dans la classification.

Cette approche permettrait de simuler les déformations des classifications au fil du temps. Elle est donc potentiellement plus prometteuse que l'approche développée par Noyons et Van Raan. Néanmoins, son applicabilité dans notre domaine n'est pas évidente, puis qu'il faut recalculer les fréquences de mots, d'auteurs, etc. à chaque nouveau document avant de reconstruire la hiérarchie des classes. Et même en se situant dans une hypothèse de vocabulaire fermé comme précédemment (ce qui serait fort restrictif), comment rendre compte visuellement de ces phénomènes ?

3 Perspectives

Pour la détection et l'analyse des évolutions thématiques dans le temps, l'intégration de techniques de classification incrémentale au sein d'une plate-forme d'analyse est une piste prometteuse. Elle suppose une évolution de l'ergonomie de l'interface d'analyse qui devra être capable de construire dynamiquement des images animées dans un espace multi-dimensionnel à partir de données stockées dans une base de données dont les valeurs évolueront au fil du temps. Ceci constitue un objectif à long terme pour la recherche sur la génération automatique d'hypertextes ergonomiques pour l'analyse de l'information.

Cette voie de recherche est une recherche appliquée de nature transversale mêlant profondément informatique, analyse de données et sciences de l'information. Elle nécessite une collaboration active avec des chercheurs de ces trois domaines. Ma formation initiale en informatique (DEA) et la réflexion que j'ai menée dans le cadre de cette recherche devraient me permettre de poursuivre cette voie en collaboration avec l'équipe Orpailleur du LORIA à Nancy, avec laquelle l'URI partage un objectif à moyen terme : construire un système de gestion de connaissances au service de la veille scientifique [POLANCO et al. 1998b].

Par ailleurs, j'entends poursuivre la réflexion que j'ai développée dans le chapitre 7 sur la constitution de bases infométriques hybrides (multi-sources, multi types de données) et notamment l'intégration de données hétérogènes. Le besoin croissant d'indicateurs européens, nationaux, régionaux, institutionnels demande, pour être satisfait, la mise en place de ces nouvelles bases de données hybrides adaptées au calcul d'indicateurs. Pour l'INIST, cela signifie la possibilité de se positionner comme un acteur important sur la scène européenne.

1. [ABITEBOUL et al. 1997] ABITEBOUL S., CLUET S., CHRISTOPHIDES V., MILO T., MOERKOTTE G., SIMEON J. - Querying Documents in Object Databases -, International Journal on Digital Libraries, 1(1), 5-19, 1997.
2. [BARRE et al. 1995] BARRE R., LAVILLE F., TEIXEIRA N., ZITT M. 'L'observatoire des sciences et des techniques : activités- définition- méthodologie' SOLARIS, 2, p.219-235, 1995.
3. [BLAIR 1988] BLAIR D.C. 'An extended relational Document Retrieval Model', Information Processing and Management Vol 24, n°3 (1988), 259-371.
4. [BORDONS et al. 1995] BORDONS M. ., ZULUETA M.A, CABRERO A . 'Identifying Research teams with bibliometric tools publications' In: Michael E.D Koenig, Abraham Bookstein (Eds), 5th International Conference of the International Society for Scientometrics and Informetrics, Learned Information Inc. Medford NJ, 83-92, 1995.
5. [BRAAM et al. 1998] BRAAM R.R., MOED H.F., VAN RAAN A.F.J., -Comparison and Combination of Co-Citation and Co-Word Clustering- , in Select Proceeding of the First International Workshop on Science and Technology Indicators, Leiden, 14-16 November 1988, p. 307-337, 1988.
6. [BALPE et al 1996] BALPE J.P, LELU A., SALEH I. ET PAPY F. - Techniques avancées pour l'hypertexte - éditions Hermès, 1996.
7. [BOUTIN et al 1998] BOUTIN E., MANNINA B., ROSTAING H., QUONIAM L. Construction automatique de réseaux : un outil pour mieux appréhender l'information provenant d'Internet, Actes JADT 98, Coord. S. Mellet, UPRESA « Bases Corpus et Langages » Université de Nice 1998.
8. [BRADFORD 1934] BRADFORD S. C. - Sources of information on specific subjects - Enginering, 137 : 85-86, Janvier 1934.
9. [BROOKES 1980] BROOKES B.C., -Information Space-, The Canadian Journal of Information Science, vol. 5, p. 199-211, 1980.
10. [BROOKES 1981] BROOKES B.C., -The Foundations of Information Science. Part IV: Information Science: The Changing Paradigm-, Journal of Information Science, vol. 3, 1981, p. 3-12
11. [CALLON et al 1983] CALLON M., COURTIAL J-P., TURNER W.A., BAUIN S. 1983 - "From Translation to Problematic Networks: An Introduction to Co-Word Analysis" in Social Science Information, vol. 22, pp. 191-235.
12. [CALLON et al 1986] M. CALLON, J. LAW, A. RIP (eds), Mapping the Dynamics of Science and Technology. London, Macmillan Press, 1986.
13. [CALLON et al 1991] M. CALLON, J-P. COURTIAL, F. LAVILLE, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry", Scientometrics, vol. 22, n° 1, p. 155-205, 1991.
14. [CALLON 1993] CALLON M. COURTIAL J.P PENAN H. - La scientométrie - Que Sais-je, PUF Paris, 1993.

15.[CAPPONI 1999] CAPPONI N. Généralisation de structures prédicatives. Application à l'analyse de l'information. Thèse de doctorat Science de l'information et de la communication, Université H. Poincarré Nancy 1, 1999.

16.[CODD 1970] CODD E. F. A relational model of data for large shared data banks, Comm. of the ACM, Vol13 (6): 377-387, 1970.

17.[COURTIAL 1990] COURTIAL J.P. - "Introduction à la scientométrie : de la bibliométrie à la veille technologique", Anthropos - Economica, Paris, 1990.

18.[DESVAL et DOU 1992] H. DESVALS, H. DOU : "La veille technologique", DUNOD, Paris 1992.

19.[DKAKI et al 1997] DKAKI T., DOUSSET B., MOTHE J. "Mining information in order to extract hidden and strategic information", Computer-Assisted Information Searching on Internet, RIAO97, pp 32-51, June 1997.

20.[DKAKI et al 1998] DKAKI T., DOUSSET B., MOTHE J. "Analyse d'informations issues du Web avec Tétralogie", VSST'98 Veille Stratégique Scientifique & Technologique, Toulouse ,Octobre 1998.

21.[DOU 1995] DOU H. Veille technologique et compétitivité, Dunod, 1995.

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

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

24.[DUCLOY et POLANCO 1992] DUCLOY J., POLANCO X. -"D'une boite à outils à la description du domaine des cognisciences", Journées d'étude ADEST "Prendre la mesure des sciences et techniques : la scientométrie en action", Paris 1-11 juin 1992.

25.[DUCLOY et al. 1991] DUCLOY J., GRIVEL L., LAMIREL J.C., POLANCO X., SCHMITT L. INIST's Experience in Hyper-Document Building from Bibliographic Databases. Proceedings of Conférence RIAO 91, Barcelone (Spain), vol 1.

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

27.[DUCOURNEAU 1998] Langages et modèles et objets, Editeurs DUCOURNEAU R. EUZENAT J. MASINI G. NAPOLI A. Collection Didactique, INRIA, 527 p., 1998

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

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

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

31.[FRANÇOIS 1998] FRANÇOIS C. - NEURODOC : un outil d’analyse de l’information -, Conférence. VSST’98 (Veille Stratégique Scientifique et Technologique), Toulouse, 19-23 octobre, 1998.

32.[GARFIELD 1972] E. Garfield, "Citation analysis as a tool in journal evaluation", Science 178, pp 471-479, 1972.

33.[GLANZEL 1996] GLÄNZEL W. ‘The Need for Standards in Bibliometric Research and Technology’, Scientometrics, vol.35, N°2, 167-176, 1996.

34.[GODIN 1995] GODIN R., MINEAU G., MISSAOUI R., MILI H. Méthodes de classification conceptuelles basées sur les treillis de Gallois et applications, Revue d’intelligence artificielle Vol. 9, n°2, pages 105-137, 1995.

35.[GOLDFARB 1990] GOLDFARB C. The SGML Handbook, Oxford, Oxford University Press. 1990.

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

37.[GRIVEL et LAMIREL 1993] GRIVEL L., LAMIREL J.C. - “An analysis tool for scientometric studies integrated in an hypermedia environment”, ICO93, 4th International Conference on Cognitive and Computer Sciences for Organizations, Montreal, (Quebec) Canada, pp146-154, 4-7 mai 1993.

38.[GRIVEL et FRANCOIS 1995a] GRIVEL L., FRANÇOIS C. "Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique", SOLARIS n° 2, Presses Universitaires de Rennes, p. 81-112, 1995. et

http//www.info.unicaen/bnum/jelec/Solaris.

39.[GRIVEL et FRANCOIS 1995b] GRIVEL L., FRANÇOIS C. Conception et développement d'un système d'information dédié à la veille scientifique, basé sur les sorties des outils de classification thématique : SDOC et NEURODOC , In : BALPE J.P, LELU A., SALEH I,.Eds, Hypertexte et hypermedia, réalisations, outils et méthodes, Paris, Editions Hermès: 109-118, 1995.

40.[GRIVEL et al. 1995] GRIVEL L., MUTSCHKE P., POLANCO X. Thematic mapping on bibliographic databases by cluster analysis: a description of the SDOC environment with SOLIS, Journal of Knowledge Organization, vol. 22, (2): 70-77, 1995.

41.[GRIVEL et al. 1997] GRIVEL L., POLANCO X., KAPLAN A. ‘A computer system for big scientometrics at the age of the World Wide Web’, Scientometrics, vol.40, N°3, 493-506, 1997.

42.[GRIVEL 1999] GRIVEL L. ‘HENOCH, un outil d’analyse de corpus d’information scientifique et technique’, Le Micro Bulletin Thématique n°3, L’information scientifique et technique et l’outil Internet, Editeur CNRS-DSI, p.27-44, 1999.

43.[GROSS 1988] GROSS G., “Structure des noms composés”, Informatique & Langue Naturelle, ILN'88, Nantes, France. Octobre 1988

44.[HABERT ET JACQUEMIN 1993] HABERT, B., JACQUEMIN C., “Noms composés, termes, dénominations complexes : problématiques linguistiques et traitement automatiques”, Traitement Automatique des Langues, 34 (2), p. 5-42,1993.

45.[HEALEY et al. 1986] P. HEALEY, H. ROTHMAN, P. HOCH, “An Experiment in Science Mapping for Research Planning”, Research Policy, vol. 15, p. 233-251, 1986.

46.[HERWIJNEN 1990] HERWIJNEN E. "Practical SGML", Kluwer Academic Publishers, 1990 .

47.[HUOT 1992] HUOT C. Analyse relationnelle pour la veille technologique : vers l’analyse automatique des bases de données, thèse de doctorat en Sciences de l’Information et Communication, Université Aix Marseille III, 1992.

48.ISO 8879 - Information processing - Text and office systems - Standard Generalised Markup Language (SGML), 155 pages, 1986.

49.[JACQUEMIN 1994] JACQUEMIN, C. - FASTR: A Unification-based Front-end to Automatic Indexing - RIAO 94 Conference Proceedings «Intelligent Multimedia Information Retrieval Systems and Management», Rockfeller University, New York, October 11-13, p. 34-47, 1994.

50.[JACQUEMIN et ROYAUTE 1994] JACQUEMIN, C., ROYAUTE J., "Retrieving Terms and their Variants in a Lexicalised Unification-Based Framework", Proceedings 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 3 - 6 July 1994, Dublin.

51.[JACOBIAK 1992] JAKOBIAK. F. Exemples commentés de veille technologique. Paris : Les Editions d'Organisation, 1992, p. 27.

52.[JACOBIAK 1996] JACOBIAK F. L’information scientifique et technique, Que Sais-je, 1996.

53.[JOUVE 1998] JOUVE 0. Sampler, manuel utilisateur, N° S5.22 /98/02/01, Compagnie des signaux, 1998

54.[KISTER et al. 1993] KISTER J., RUAU O., QUONIAM L., DOU H. Application des outils bibliométriques en chimie analytique 4 ème Journées sur l'information élaborée Ile Rousse, Revue Française de bibliométrie 12, p. 437-456, 1993.

55.[KOHONEN et al. 1995] KOHONEN T. KASKI S. LAGUS K. HONKELA T. - Very large two level SOM for the browing of newsgroups - 5th International WWW Conference Paris 1995.

56.[KOPCSA et SCHIEBEL 1998] KOPCSA A., SCHIEBEL E. - Science and technology mapping : a new iteration model for representing relationships - Jasis

49 (1) :7-17 1998.

57.[KRUSKAL 1964] KRUSKAL J.B. - Multidimensonal scaling by optimizing goodness of fit to a nonmetric hypothesis - Psychometrika, 29 :1-28, 1964.

58.[LAFOUGE 1998] LAFOUGE T. Mathématiques du document et de l'information. Bibliométrie distributionnelle, Habilitation à diriger des recherches, RECODOC, Univ. Lyon 1, Oct. 1998

59.[LEBART et SALEM 1988] LEBART L. SALEM A. - Analyse statistique des données textuelles -, DUNOD, Paris 1988, 207 pages.

60.[LECOADIC 1994] LECOADIC Y. - La science de l'information - Que Sais-je, PUF Paris, 1994.

61.[LELU 1990] LELU A. - "Modèles neuronaux pour données textuelles - Vers l'analyse dynamique des données" - Journées ASU de statistiques, Tours, France.

62.[LELU 1990] LELU A. - "Modèles neuronaux de projection associative et analyse des données" - Approches symboliques et numériques pour l'apprentissage de connaissances à partir des données - sous la direction d'E. DIDAY et Y. KODRATOFF, pp 283-305, CEPADUES, Toulouse, 1990.

63.[LELU et FRANCOIS 1992] LELU A. et FRANCOIS C. - "Automatic generation of hypertext links in information retrieval systems", communication au colloque ECHT'92, Milan, D. Lucarella & al. eds, ACM Press, New York.

64.[LELU 1993] LELU A. - "Modèles neuronaux pour l'analyse de données documentaires et textuelles" Thèse de doctorat de l'université de Paris VI. 4 mars 1993, 238 pages. -

65.[LELU et al 1997] LELU A. , Tisseau-Pirot A.G., Adnani A.'Cartographie de corpus textuels évolutifs : un outi pour l'analyse et la navigation' Hypertextes et Hypermedia, Vol1. N°1, éditions Hermès, Paris, 1997

66.[LELU et al 1998] LELU A., HALLEB M., DELPRAT B. 'Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-Grammes » Actes des 4è journées internationales d'analyse statistique des données textuelles, Nice 1998.

67.[LEVEILLE 1998] LEVEILLE V., ROSTAING H., QUONIAM L. Création d'hypertextes automatiques appliqués à la veille, VSST'98 Veille Stratégique Scientifique & Technologique, Toulouse ,Octobre 1998.

68.[LEVY 1990] LEVY P. 'Les technologies de l'intelligence, Collection Points Sciences, Edition La découverte, 234p, 1990.

69.[LOTKA 1927] LOTKA A.J. The frequency distribution of scientific productivity. Journal of the Washington academy of sciences, 16(2) :317-323, Juin, 1927.

70.[MARCOTORCHINO 1991] MARCOTORCHINO J.F. , seriation problems : an overview, Applied stochastics Models and Data Analysis, Vol. 7 N°2, 1991.

71.[MARTEAU 1995] MARTEAU P.F., KRUMEICH C. Analyse sémantique pour le veille technologique, IDT. Information, documentation, transfert des connaissances, Paris France; Pp. 258-263, 1995;

72.[MEINKE et ATHERTON 1976] P. MEINCKE AND P. ATHERTON,

"Knowledge Space: A Conceptual Basis for the Organization of Knowledge", Journal of the American Society for Information Science, vol. 27, p. 18-24, 1976.

73.[MICHARD 1998] MICHARD A. 'XML Langage et application' Editions Eyrolles, 361 p, 1998

74.[MICHELET 1988] MICHELET B. L'analyse des associations. Thèse de doctorat en Sciences de l'information, Université de Paris VII, 1988.

75.[MOED 88] MOED H.F 'The use of On-line databases for bibliometric analysis', In L. Egghe and R. Rousseau (editors), Informetrics 87/88 (Elsevier Science Publishers), Amsterdam), 145-158, 1998.

76.[MOED 95] MOED H.F, DE BRUIN R.E, Van LEEUWEN TH. 'New bibliometric tools for the assessment of National Research Performance : Database description, overview of indicators and first applications', Scientometrics, Vol.33, n°3, 381-422, 1995.

77.[MOED 95b] MOED H.F, Van LEEUWEN TH. 'Improving th accuracy of the ISI's journal impact factor, Journal of the American Society for Information Science, 46, 381-422, 1995.

78.[MOED 1996] MOED H.F. 'Differences in the construction of SCI Based Bibliometric Indicators among Various Producer : A first Overview' , Scientometrics, , vol.35, N°2, 177-192, 1996.

79.[NEDERHOF et al. 1989] A.J. NEDERHOF, R.A. ZWAAN, R.E. DE BRUIN, P.J. DEKKER, "Assessing the Usefulness of Bibliometric Indicator for the Humanities and the Social and Behavioural Sciences: A Comparative Study", Scientometrics, vol. 15, n° 5-6, p. 423-433, 1989.

80.[NAUER 99] NAUER E. 'De l'importance de la normalisation en bibliométrie', Journées d'études sur les systèmes d'information élaborée de la SFBA, Ile Rousse, Corse, 27 septembre-1er octobre 1999

81.[NOYONS et VAN RAAN 1998] Noyons E., Van Raan A. Monitoring scientific developments from a dynamic perspective Jasis 49 (1) :68-81 1998.

82.[PETERS et VAN RAAN 1993] PETERS H.P.F., VAN RAAN A.F.J. - "Co-word based science maps of chemical engineering, Part II : Representations by combined clustering and multidimensional scaling" Research Policy, vol.22, 1993, p.47-70.

83.[POLANCO et al. 1993] POLANCO, X., L. GRIVEL, C. FRANÇOIS ET D. BESAGNI, "L'infométrie, un programme de recherche", Journées d'études "Les systèmes d'information élaborée". Ile Rousse, Corse, France, 9-11 Juin1993, texte n° 3.

84.[POLANCO 1993] POLANCO, X. , "Analyse de l'information scientifique et technique. Construction de clusters de mots-clés", Sciences de la société, n° 29, p. 111-126.

85.[POLANCO et FRANCOIS 1994] POLANCO X., FRANCOIS C. - "Les enjeux de l'information scientifique et technique à travers une analyse d'infométrie cognitive utilisant une méthode de classification automatique et de représentation

conceptuelle (NEURODOC)", Actes du colloque ORSTOM/UNESCO "Les sciences hors occident au XXè siècle, Paris 19-23 septembre 1994.

86.[POLANCO et GRIVEL 1995] POLANCO X., GRIVEL L. - "Mapping knowledge: the use of co-word analysis techniques for mapping a sociology data file of four publishing countries (France, Germany, United Kingdom and United State of America), Internation. journal of Scientometrics and Informetrics, Vol1 (2),pp123-137, 1995.

87.[POLANCO 1995] POLANCO X. 'Aux sources de la scientométrie', SOLARIS, Vol 2 «Les sciences de l'information : bibliométrie, scientométrie, infométrie, sous la direction de Jean-Max Noyer ». Edition : Presses Universitaires de Rennes, pp.13-78, 1995.

88.[POLANCO et al. 1995] POLANCO X., GRIVEL L., ROYAUTE J. How to do things with terms in informetrics: terminological variation and stabilization as science watch indicators, In: Michael E.D Koenig, Abraham Bookstein (Eds), 5th International Conference of the International Society for Scientometrics and Informetrics, Learned Information Inc. Medford NJ: 435-444, 1995.

89.[POLANCO et al. 1997] POLANCO X., FRANÇOIS C., KEIM J.P. Artificial Neural Network Technology for the classification and Cartography of Scientific and Technical Information, to be published in Proceedings 6th International Conference of the International Society for Scientometrics and Informetrics, Jerusalem, June 16-19 1997.

90.[POLANCO 1997] POLANCO X. -La notion d'analyse de l'information dans le domaine de l'information scientifique et technique -, Colloque INRA, 21-23 octobre 1996, Tours. P. Volland-Neil, coord. L'information scientifique et technique : Nouveaux enjeux documentaires et éditoriaux ; Paris, INRA, 1997, pp. 165-172.

91.[POLANCO et al. 1998] POLANCO X., FRANÇOIS C., OULD LOULY A. « For Visualization-Based Analysis Tools in Knowledge Discovery Process: A Multilayer Perceptron versus Principal Components Analysis - A Comparative Study », J.M. Zytkow and M. Quafafou (eds) *Principles of Data Mining and Knowledge Discovery*. Second European Symposium, PKDD'98, Nantes, France, 23-26 September 1998. Lecture Note in Artificial Intelligence 1510. Subseries of Lecture Notes in Computer Science. Berlin, Springer, pp. 28-37, 1998.

92.[POLANCO et al. 1998b] POLANCO X., FRANÇOIS C,. ROYAUTE J., GRIVEL L., BESAGNI D., DEJEAN M., OTTO C. « Organisation et gestion des connaissances en veille scientifique et technologique», VSST'98 (Veille Stratégique Scientifique et Technologique), Toulouse, 19-23 octobre, Actes éditées par l'Université Paul Sabatier, p.328-337, 1998.

93.[POPPER 1979] K.P. POPPER, Objective Knowledge. Oxford: The Clarendon Press, 1979.

94.[PRICE 1965] D. de S. PRICE, "Network of Scientific Papers", Science, vol. 149, n° 3683, 1965, p.510-515.

95.[PRICE 1986] D. de S. PRICE, "The Citation Cycle", p. 269 in Little Science, Big Science ... and Beyond. New York, Columbia University Press, 1986.

96.[PRICE 1984] D. de S. PRICE, “The Science-Technology Relationship, the Craft of Experimental Science, and Policy for the improvement of High Technology Innovation”, Research Policy, vol. 13, 1984, p. 3-20.

97.[QUONIAM L. 1988] QUONIAM L. ‘Bibliométrie Informatisée et Information Stratégique’, Thèse de doctorat. en Sciences de l'information et de la communication. Université Aix-Marseille III.. pp. 330, 1988.

98.[QUONIAM L. 1992] QUONIAM L. ‘Bibliométrie sur références bibliographiques: méthodologie’ in: La Veille Technologique: l'Information scientifique, technique, industrielle. DUNOD, 1992.

99.[Rapport Inria N° 3198] - MULLER C., POLANCO X ., ROYAUTE J. TOUSSAINT Y. - Acquisition et structuration des connaissances en corpus : éléments méthodologiques Rapport Inria N° 3198, 1997.

100.[ROSTAING 1996] ROSTAING H. ‘La bibliométrie et ses techniques’, Edition : sciences de la société, coll : « Outils et méthodes », 131p. 1996.

101.[ROUSSEAU 1998] ROUSSEAU F. - L’analyse de corpus d’information comme support de la veille stratégique - Document numérique (2), 177-202, juin 1998 .

102.[ROYAUTE et JACQUEMIN 1993] ROYAUTE J., JACQUEMIN C., "Indexation automatique et recherche de noms composés sous leurs différentes variations". Informatique & Langue Naturelle, ILN'93, Nantes, France, 1993.

103.[ROYAUTE 1999] ROYAUTE J. Les groupes nominaux complexes et leurs propriétés : application à l’analyse de l’information, Thèse de doctorat en informatique, Université H. Poincaré Nancy I, 1999.

104.[SERRES 1995] SERRES A. L’hypertexte, une histoire à revisiter, Documentaliste, vol 32 n°2, 71-83, 1995.

105.[SALTON 1971] G. SALTON : "The SMART retrieval system - Experiments in automatic document processing", Englewoods Cliff, New Jersey, Prentice Hall Inc., 1971.

106.[SALTON 1989] G. SALTON : "Automatic text processing : the transformation, analysis and retrieval of information by computer, New York, Addison Wesley, 1989.

107.[SMALL et GARFIELD 1988] H. SMALL, E. GARFIELD,"The Geography of Science: Disciplinary and National Mappings", in Science Citation Index 1988, Philadelphia: Institut for Scientific Information, p. 46-58, 1988.

108.[SMALL 1973] SMALL H., "Co-citation in the scientific litterature: A new measure of the relationship between two documents", Journal of the American Society of Information Science 24, pp. 265-269, 1973.

109.[SMALL 1995] SMALL H. Relational bibliometrics, In: Michael E.D Koenig, Abraham Bookstein (Eds), 5th International Conference of the International Society for Scientometrics and Informetrics, Learned Information Inc. Medford NJ: 525-530, 1995.

110.[SMALL 1997] SMALL H. - Update on science mapping : creating large document spaces Scientometrics - 38 (2) : 275-293, 1997.

111.[SMALL 1999] SMALL H. - Visualizing science by citation mapping- Jasis 50

(9) :799-813, 1999.

112.[TEIL 1991] TEIL G. ‘Candide, un outil de sociologie assistée par ordinateur’, Thèse de doctorat du Centre de Sociologie et Innovation Ecole des Mines de Paris, 1991.

113.[TURNER et al. 1998] W. TURNER, G. CHARTON, F. LAVILLE, B. MICHELET, “Packinging Information for Peer review: New Co-word Analysis Techniques”, in A.F.J. van Raan (ed), Handbook of Quantitative Studies of Science and Technology. Amsterdam: Elsevier Science Publisher, 1988, p. 291-323.

114.[TURNET 1994] TURNER W. - “Penser l’entrelacement de l’Humain et du Technique : les réseaux hybrides d’intelligence “- Solaris n°1 “Pour une nouvelle économie du savoir”, Presses universitaires de Rennes, p.21-50, 1994.

115.[VINKLER 96] VINKLER P. ‘Standardization of Scientometric Indicators’, vol.35, N°2 (1996), 237-245.

116.[WINSTON 1977] P. H. WINSTON, Artificial Intelligence. London: Addison Wesley Publishing Co., 1977.

117.[WOLFRAM 1996] WOLFRAM D. Inter-Record linkage structure in a hypertext bibliographic retrieval system Jasis 46 (10):765-774, 1996.

118.[ZIPF 1949] ZIPF G.K. - Human Behavior and the Principle of Least Effort - Addison-Wesley, 1949.

119.[ZITT et BASSECOULARD 1994] ZITT M. , BASSECOULARD E. Development of a method for detection and trend analysis of research fronts built lexical or cocitation analysis, Scientometrics, Vol.30, (1): 333-351, 1994..

120.[ZITT et BASSECOULARD 1996] ZITT M. , BASSECOULARD E. Reassessment of co-citation methods for science indicators: effects of methods improving recall rates, Scientometrics, Vol.37, (2): 223-244, 1996.

121.[ZITT 1996] ZITT M. , TEIXEIRA N. ‘Science Macro-Indicators : some aspects of OST Experience Scientometrics’, vol.35, (2 : 209-222, 1996.

Annexe 1[1]

L'INFOMETRIE, UN PROGRAMME DE RECHERCHE

Cet article est l'article fondateur du Programme de Recherche Infométrie, 'ancêtre' de l'Unité Recherche et Innovation où j'ai effectué cette thèse. Il développe en particulier les objectifs et les réalisations principales de ce programme de recherche en 1993.

---

1. [1] *POLANCO X., GRIVEL L., FRANÇOIS C., BESAGNI D. "L'infométrie, un programme de recherche", Journées d'études sur les systèmes d'information élaborée de la SFBA, Ile Rousse, Corse, Document n° 3 des Actes, 9p, 1993.*

1. Introduction.-

La mission de ce programme est le développement d'une recherche appliquée dont le but principal est de fournir à l'INIST des outils d'analyse de l'information scientifique et technique (IST).

Les techniques infométriques et les bases de données dont elles sont issues doivent être considérées comme un dispositif de représentation ou de visualisation de l'état de la connaissance scientifique et de la pratique de ses acteurs.

Nos travaux doivent permettre de répondre à une demande sans cesse croissante en information "élaborée" de la part de chercheurs, mais aussi des responsables de l'industrie et de la recherche, ainsi que des équipes qui analysent les activités de recherche (comme le montre par ailleurs l'ouvrage sous la direction de Hélène Desvals et Henri Dou, *La veille technologique.* Paris, DUNOD, 1992)

2. Définitions.-

On entend par infométrie l'ensemble d'activités métriques concernant le domaine particulier de l'information scientifique et technique (IST) :

a) *Bibliométrie* : celle-ci a été définie en 1969 comme "l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication" (A. Pritchard, "Statistical Bibliography or Bibliometrics?", *Journal of Documentation,* vol. 25, n° 4, December 1969, p. 348-349 ; voir aussi R. N. Boadus, "Toward a Definition of «Bibliometrics», *Scientometrics,* vol. 12, n° 5-6, 1987, p. 373-379).

b) *Scientométrie* : on peut la considérer comme la bibliométrie spécialisée au domaine de l'IST. Toutefois, la scientométrie désigne d'une manière générale l'application de méthodes statistiques à des données quantitatives (économiques, humaines, bibliographiques) caractéristiques de l'état de la science. Ainsi par exemple dans la présentation de la revue *Scientometrics,* on peut lire que la scientométrie comprend la "research concerned with quantitative features and characteristics of science", ainsi que "the development and the mechanism of science (...) studied by means of (statistical) mathematical methods".

c) *Infométrie* : terme adopté en 1987 par la F.I.D. pour désigner l'ensemble des activités métriques relatives à l'information, couvrant aussi bien la bibliométrie que la scientométrie (voir L. Egghe et R. Rousseau, éds., *Informetrics 87/88.* Amsterdam, Elsevier, 1988, p. IV ; voir aussi dans ce même ouvrage la référence que fait dans ce sens B. C. Brookes dans son article "Comments on the Scope of Bibliometrics", p. 29).

3. Les fonctions de l'infométrie.-

Les fonctions de l'infométrie sont l'analyse, l'évaluation et la représentation graphique de l'IST au moyen des méthodes statistiques, mathématiques et d'analyse de données ; nous nous proposons également d'explorer l'application de techniques

(«non-quantitatives») comme celles qui sont générées dans les domaines de l'intelligence artificielle et des systèmes experts (voir par exemple R. Davis, éd., *Intelligent Information Systems.* Chichester, U.K., Ellis Horwood Limited & John Wiley & Sons, 1986).

Les fonctions de l'infométrie sont donc l'analyse, l'évaluation et la représentation graphique de l'IST. Au prix d'une analyse un peu sommaire, nous les définissons de la manière suivante :

a) *L'analyse* a pour objectif de répondre à des questions d'ordre stratégique et de veille scientifique ou technologique. Il s'agit de produire une "information de l'information".

b) *L'évaluation* de l'IST est de deux types, l'un est l'évaluation métrique des flux d'information ; l'autre est l'évaluation de qualité de l'information traitée.

c) *La représentation graphique* de l'IST (ou infographie) est l'élaboration de cartes où l'on peut positionner les contenus de l'information, mais aussi les acteurs de la recherche (auteurs, institutions, pays). Le but est ici de fournir une représentation de la structure de l'information à un moment donné de son développement sur un espace à deux ou trois dimensions.

4. Des techniques d'analyse.-

Les techniques que nous avons développées sous UNIX sont 1) la *méthode des mots associés* (à partir de la thèse de doctorat de B. Michelet, *L'analyse des associations.* Université de Paris 7. Paris, 1988) et 2) la *méthode de k-means axiales* (à partir d'un modèle défini par A. Lelu, "Modèles neuronaux pour données textuelles", *Journées ASU de Statistique*, Tours, 25 mai - 1er juin 1990) : ce sont respectivement les programmes NEURODOC et SDOC.

Ces outils permettent de structurer l'information puis de la traiter sous la forme d'un hypertexte. Pour le moment, limitons nous au fait qu'ils constituent des moyens d'organiser thématiquement l'information.

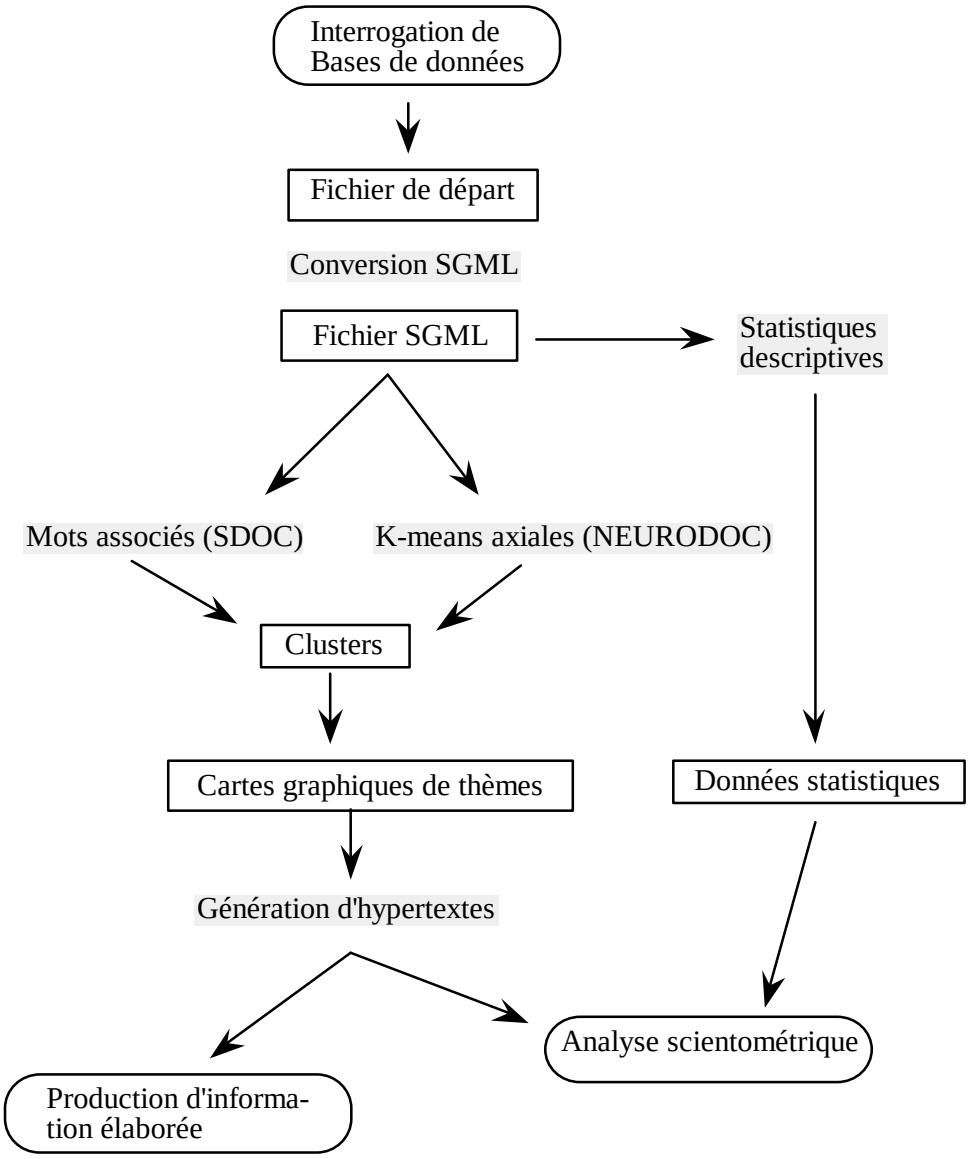
Ainsi, au lieu de parcourir une somme d'information dans un ordre séquentiel, une simple liste de références, une succession de notices bibliographiques, nous avons ici une façon de suivre un ordre thématique qui est construit à partir des données bibliographiques elles-mêmes.

Dans un fichier comportant des données bibliographiques, chaque article scientifique apparaît comme une unité qui se juxtapose à d'autres. Or, la connaissance scientifique n'est pas faite d'une juxtaposition d'éléments mais constitue une réseau d'associations multiples entre concepts, techniques, appareillages, théories, domaines d'application, méthodes, etc. On sait, par exemple, qu'il existe des thèmes de recherche autour desquels se mobilisent les intérêts des chercheurs d'un domaine particulier. Dans la mesure où chacun de ces thèmes est indiqué par des mots-clés, on peut supposer que l'association entre les termes d'indexation rende visible la trame d'un domaine de recherche.

L’avantage de l’utilisation de ces méthodes «scientométriques» est que l’on ne passe pas par un code de classement préalablement établi et figé. On suit le développement de la recherche et ses agencements tels qu’ils se présentent au niveau de la littérature scientifique, tout en sachant qu’il existe toujours le problème de l’indexation, à savoir la relation de pertinence entre les mots-clés utilisés (vocabulaire d'indexation) et le «contenu objectif» du texte scientifique.

On a observé que l'information est produite en abondance et stockée systématiquement, mais qu’elle n’est pas utilisée efficacement. Les atomes d'information sont des fragments inutilisables à moins qu'ils ne soient convenablement adaptés aux structures de connaissance de ceux qui les demandent ou les cherchent. De là cette idée qu’il faut présenter l'information que l'on offre dans le contexte d'une structure cognitive pertinente ou appropriée, de façon à ce que l'usager puisse ainsi percevoir sa pertinence ou sa garantie informationnelle.

5. Schéma infométrique.-



Ce schéma synthétise la démarche que nous avons mis en place ; cette chaîne infométrique obéit au principe de la modularité, c'est-à-dire que chaque opération constitue un module informatique ; l'ensemble de ces modules est à la disposition de l'utilisateur dans une bibliothèque dénommée ILIb. Cette bibliothèque constitue une véritable boîte à outils pour le traitement de l'IST, dans ce cas précis, pour l'analyse scientométrique et la production d'une information élaborée intéressant au premier chef l'analyse stratégique ainsi que la veille scientifique et l'aide à la décision (aussi bien dans la gestion de l'IST que dans la politique scientifique).

6. Les éléments de la chaîne infométrique.-

1 - Fichiers et Formats
2 - Indexation
3 - Bibliométrie
4 - Outils scientométriques
5 - Infographie
6 - Hypertexte
7 - Edition

En signalant ces différents éléments, nous voulons mettre en valeur notre approche informatique fondée sur la modularité par décomposition en programmes qui s'échangent des flux d'information (telle qu'elle est exposée en J. Ducloy, P. Charpentier, C. François, L. Grivel, "Une boîte à outils pour le traitement de l'Information Scientifique et Technique", *Actes des 4es. Journées Internationales Le Génie logiciel et ses applications.* Toulouse, 9-13 Décembre 1991, p. 239-254). Ces programmes sont développés sous UNIX, un système multi-utilisateur et multi-tâche d'exploitation et de développement, qui grâce à son mécanisme de "pipe" permet de combiner différents outils.

1 - Fichiers et Formats

La première étape est l'élaboration des fichiers à partir desquels une analyse se fera, Ce qui implique un travail sur les formats des notices bibliographiques afin de les rendre exploitables tout au long de la chaîne. C'est donc la définition d'un format pivot unique.

L'idée fondamentale est l'utilisation du balisage de la norme SGML (Standard Generalized Markup Language) pour décrire toutes les données quelle que soit leur organisation. Une fois que toutes les données sont homogénéisées dans un format pivot unique, il est plus facile de concevoir des outils génériques utilisant les propriétés du balisage SGML.
Voir C. François, *Analyse de références bibliographiques conformes à la norme ISO 2709 et conversion vers la norme SGML.* Rapport de stage DESS Informatique, INIST-CNRS, ISIAL, Université de Nancy 1,1990 ; N. Dusoulier et J. Ducloy, "Processing of data and exchange of records in scientific and technical information center. Formats: what for?. Communication à CCF-UNIMARC Workshop, Florence, 5-6 juin 1991.

2 - Indexation

Les programmes NEURODOC et SDOC ont comme «input» des mots-clés, qui sont des indicateurs du contenu des articles scientifiques.

Ces mots-clés peuvent être fournis par les notices elles-mêmes, ce qui pose le problème de leur adéquation aux besoins de ls scientométrie. Ceci implique de nous doter d'un outil d'indexation assistée par ordinateur.

Voir à ce sujet J. Royauté, L. Schmitt et E. Olivetan, "Les expériences d'indexation à l'INIST". *Actes du 15e Colloque International en Linguistique Informatique* : COLING-92, Nantes, 23-28 août 1992, vol. III, p.1058-1063.

NB : Nous travaillons à ce sujet en collaboration étroite avec le programme de recherche INDEXATION qui a pour mission, sous la responsabilité de L. Schmitt, de doter à l'INIST d'outils d'aide à l'indexation.

3 - Bibliométrie

Cette étape correspond à l'application d'outils statistiques pour analyser notamment la distribution et la fréquence des données bibliographiques. L'objectif est de caractériser, à l'aide de tableaux et de graphes, la littérature scientifique dans un domaine déterminé.

On peut ainsi quantifier sa magnitude (nombre d'articles, nombre de revues), son actualité (selon la date de publication), sa localisation (selon le pays d'édition des revues scientifiques), l'importance des périodiques scientifiques (selon le nombre d'articles dont ils sont la source au cours d'une période déterminée) et la localisation des auteurs (selon leur appartenance institutionnelle).

Le traitement statistique se fait en trois étapes, utilisant un ensemble de programmes d'analyse statistique descriptive permettant de créer un fichier résultat directement exploitable en sortie papier ou sous un tableur de type EXCEL par exemple.

Première étape : création de fichiers inverses.

Seconde étape : analyse statistique des notices : 1) comptage de références ; 2) distribution par langue ; 3) distribution par pays d'affiliation des auteurs ; 4) distribution par type de document ; 5) distribution par date de publication ; 6) distribution par titres de périodiques ; 7) distribution par pays d'édition de ces titres.

Troisième étape : analyse statistique des mots-clés : 1) nombre de mots-clés ; 2) moyenne par notices ; 3) distribution des mots-clés par notices ; 4) distribution des mots-clés par fréquence.

Dans cette phase de la chaîne infométrique, on utilise des shell-script UNIX encapsulant des programmes écrits en langage C. On peut envisager l'utilisations des outils disponibles sur le marché, comme par exemple MATLAB en mathématiques,

SAS en statistiques et SPAD.N en analyse de données, et le développement d'un génie mathématique plus sophistiqué.

4 - Outils scientométriques

Ce sont les programmes NEURODOC et SDOC (voir leur fiche technique dans la section 7 de cette communication).

NEURODOC est un ensemble de modules implémentant la méthode de K-means axiales.

Voir C. François, N. Appel, G. Bloch, M. Gabsi, J. Ducloy, "NEURODOC, Nouveaux profils documentaire", *Compte rendu de fin d'étude d'une recherche financée par le Ministère de la Recherche et de la Technologie*, décembre 1991; A. Lelu et C. François, "Information retrieval based on a neural unsupervised extraction of thematic fuzzy clusters", communication à *Neuro-Nîmes 92 : Les réseaux neuro-mimétiques et leurs applications*, Nîmes, 2-6 novembre 1992 ; A. Lelu et C. François, "Hypertext paradigm in the field of information retrieval: a neural approach", communication à *Fourth ACM conference on Hypertext : ECHT'92*, Milan (Italie), 30 novembre- 4 décembre 1992.

SDOC est un ensemble de modules implémentant la méthode des mots associés.

Voir X. Polanco, L. Schmitt, D. Besagni, L. Grivel, "A la recherche de la diversité perdue : est-il possible de mettre en évidence les éléments hétérogènes d'un front de recherche?", *Actes des journées d'étude sur Les systèmes d'information élaborées*, organisées par SFBA, Ile Rousse, Corse, 6-5 juin 1991, p. 273-292 ; J. Ducloy et X. Polanco, "D'une boîte à outils à la description du domaine des Cognisciences", *Actes des journées d'étude La scientométrie en action*, organisées par l'ADEST, Paris, 1-2 juin 1992, p.65-73.

5 - Infographie

C'est la représentation graphique des résultats, dans notre cas, l'élaboration de cartes comme sortie de NEURODOC et de SDOC. Les cartes NEURODOC sont construites à l'aide d'une Analyse en Composantes Principales (ACP) et affichées en Hypercard ; les cartes SDOC s'affichent sous UNIX au moyen de trois modules graphiques développés en langage C et utilisant le système de composition de documents LATEX (voir E. Nataf, *Composition de page en LATEX - Création d'outils graphiques pour la scientométrie.* Rapport de stage. INIST / I.U.T. de l'Université de Nancy 2, Option génie informatique. 1992).

6 - Hypertexte

La génération d'hypertextes à l'aide des logiciels Hypercard, Folio sur PC ou Zen[2] sous UNIX est à considérer selon un double objectif : 1) fournir à l'utilisateur final un

[2] Protoype d'éditeur hypertexte développé par Bull-Cediag dans le cadre du projet européenn KWICK

document hypertexte qui lui permet de naviguer dans un ensemble de références bibliographiques à partir d'une carte globale des thèmes ; 2) fournir un instrument de travail pour l'analyse scientométrique.

Voir J. Ducloy, L. Grivel, J-Ch. Lamirel, X. Polanco, L. Schmitt, "INIST's Experience in Hyper-Document Building from Bibliographic Data Bases". Communication à la Conférence RIAO 91- Barcelone (Spain), 2-5 Avril 1991; J. Ducloy et A. Lelu, "NEURODOC : construction d'hyperdocuments à l'aide de procédés neuronaux". Communication à Génie Linguistique 91, Versailles, 16-17 janvier 1991 ; L. Grivel et J-Ch. Lamirel, "SDOC, A Generator of Hypertext Structures". Communication à 2th. Coference Multimedia Information - Cambridge (UK), 15-18 july 1991).

7 - Edition

Deux voies sont possibles pour l'édition sur support papier : 1) utiliser des compooseur tels que Troff ou LATEX sous UNIX, et 2) travailler dans des environnement plus sophistiqués tels que celui du logiciel INTERLEAF qui permet de créer et d'éditer des documents structurés.

7. Fiche technique des programmes

1 —Technique statistique
2 —Représentation des données documentaire
3 —Méthode de classification
4 —Anatomie d'une classe
5 —Nom du thème
6 —Documents associés aux thèmes
7 —Les paramètres à fixer
8 —Position des thèmes sur un plan (cartes)

A) Programme **NEURODOC**

1 - Technique statistique
K-means axiales

2 - Représentation des données documentaire
Matrice creuse : documents / descripteurs
où : 1 dans case (i,j) si document i indexé par descripteur j, 0 sinon.
Les descripteurs sont les axes du repère où les documents sont positionnés. Ils jouent tous le même rôle et sont perpendiculaires 2 à 2.
Le repère utilisé est donc de type "euclidien".

3 - Méthode de classification
Classification non hiérarchique des documents dans l'espace défini par les mots-clé, puis projection simultanées des documents et descripteurs sur les axes représentant les classes. Cf. méthode des K-means axiales.

4 - Anatomie d'une classe
Une classe est un démi-axe défini dans l'espace des mots-clés, passant par l'origine de cet espace. Sur cet axe s'ordonnent les documents et mots-clés. Une classe est donc constituée de deux listes de mots-clés et de documents triés par ordre de "pertinence" décroissante par rapport au type de la classe.
5 - Nom du thème
Le mot-clé de poids le plus fort sur l'axe représentant le thème est utilisé comme nom du thème. Cette heuristique est très frustre, la révision de ce nom par un expert du domaine semble nécessaire.

6 - Documents associés aux thèmes
Les documents ayant une projection, sur l'axe représentant le thème, supérieure à un seuil (paramètre de la méthode).

7 - Les paramètres à fixer
-nombre de classes
-mode d'initialisation des classes
-seuil des documents
-seuil des descripteurs
-nombre maximum de documents par classe
-nombre maximum de descripteurs par classe.

8 - Position des thèmes sur un plan
ACP des thèmes obtenus définis dans l'espace des mots-clés.

B) Programme **SDOC**

1 - Technique statistique
Méthode du simple lien

2 - Représentation des données documentaires
Matrice creuse : (documents / descripteurs)
où : 1 dans case (i,j) si document i indexé par descripteur j 0 sinon.
Cette matrice permet de définir la co-occurence entre 2 mots-clés, puis un coefficient d'association entre ceux-ci. Un réseau d'associations entre les mots-clés est donc défini.
Ces associations définissent une "distance" entre les mots-clés.

3 - Méthode de classification
Classification hiérarchique des mots-clés basée sur la" distance" définie ci-dessus.Cf. méthode du simple lien.

4 - Anatomie d'une classe
Une classe est un sous-ensemble du réseau des mots-clés. Elle est donc constituée :
-d'une liste de mots-clés internes
-d'une liste d'associations internes
-d'une liste d'associations externes
-d'une liste de mots-clés externes

5 - Nom du thème
Le mot-clé appartenant à la liste de mots-clés internes figurant dans le plus grand nombre d'associations (internes et externes) est utilisé comme nom du thème.

6 - Documents associés aux thèmes
Les documents ayant au moins deux mots-clés appartenant à la liste de mots-clés internes, ou un mot-clés appartenant à la liste de mots-clés internes et un mot-clés appartenant à la liste de mots-clés externes.

7 - Les paramètres à fixer
-mode de calcul des coefficients d'associations
-taille de classes
-nombre max d'associations internes
-nombre max d'associations externes
-nombre maximum de documents par classe

8 - Position des thèmes sur un plan
Axe horizontal (X) : associations externes
Axe vertical (Y) : associations internes

8. La connaissance objective.-

Nous exposons ici les bases théoriques de notre programme infométrique que nous entendons circonscrire prioritairement au domaine des sciences de l'information.

Selon le philosophe des sciences Karl Popper, il existe le monde des phénomènes physiques et sociaux, le monde subjectif des états de conscience, des états mentaux ou des dispositions comportementales, celui du sujet connaissant, et par rapport auquel la connaissance écrite, celle qui est véhiculée par la littérature scientifique et que nous analysons représentent la «connaissance objective».

Ceci induit deux catégories de problèmes concernant l'étude de la connaissance : la première comprend les problèmes relatifs aux actes de production ou de formation de connaissance ; la seconde comprend les problèmes relatifs aux structures de la connaissance produite, au sens objectif d'écrite et publique.

C'est cette deuxième catégorie de problèmes qui constitue l'objet de notre travail. Il s'agit d'analyser le «contenu» de la connaissance produite, afin de pouvoir fournir une représentation de sa structure à un moment donné de son développement.

Il est important de ne pas négliger la rétroaction des produits de la recherche sur le comportement des producteurs (chercheurs). L'autonomie de l'IST et sa rétroaction sur le monde de la recherche, de l'enseignement ou de l'industrie, sont un fait important du développement de la connaissance scientifique et technique.

Au sujet de la notion, de "connaissance objective", voir l'article de G. Frege, "Sens et dénotation" (1892), dans ses *Ecrits logiques et philosophiques*, Paris, Editions du Seuil, 1971, pp.102-126; voir surtout l'essai de K. Popper, "Une épistémologie sans

sujet connaissant" (1967), dans *La connaissance objective.* Paris : Aubier, 1991, ch.3, pp.177-242; quant à son application dans le domaine des sciences de l'information, voir B.C. Brookes, "The Foundations of Information Science" (1980-81), in *Journal of Information Science,* vol. 2 (1980), pp. 125-133 (Part I); pp. 209-221(Part II) et pp. 269-275 (Part III); vol. 3 (1981), pp. 3-12 (Part IV). Selon Brookes (Part I, p. 127) : "What information science needs at its roots, it seem to me, is an objective rather a subjective theory of knowledge".

9. Information et Connaissance.-

La relation entre information et connaissance est exprimée par «l'équation de Brookes» :

$$C[S] + \Delta I = C[S + \Delta S]$$

selon laquelle la structure de connaissance C[S] se transforme dans la structure C[S + ΔS] par l'apport d'information ΔI ; ΔS indique l'effet de la modification (Voir B.C. Brookes, "The Foundations of Information Science" (1980-81), *Journal of Information Science,* vol. 2, 1980, p. 131).

L'équation a une forme pseudo-mathématique, mais sous cette forme elle sert à souligner le peu que nous savons sur les modes selon lesquels notre connaissance croît.

Si les termes de l'équation étaient mesurables, ils devraient l'être selon la même mesure, autrement dit *l'information* et *la connaissance* appartiennent à la même classe ou espèce ; mais il convient de ne pas substituer ΔI par ΔC dans l'équation pour la simple raison qu'un même apport d'information ΔI, peut avoir des effets différents sur des structures de connaissance différentes.

L'information diffère des données de la sensibilité (*sense-data*), par le fait que ces dernières doivent être interprétées subjectivement par une structure de connaissance afin de devenir de l'information.

Plus important encore, «l'équation de Brookes» suppose que la croissance de la connaissance n'est pas simplement additive. L'absorption d'information dans la structure de connaissance peut provoquer non pas simplement une addition mais un certain réajustement de la structure, tel qu'un changement dans les liens entre deux ou plusieurs concepts admis.

10. Documents et Information.-

A l'aide de deux indicateurs, le périodique et l'article scientifique, on a mesuré la science, et l'on a pu dégager une «loi de croissance exponentielle». Mais aucune croissance ne peut rester exponentielle indéfiniment. Il a donc fallu envisager l'hypothèse de la nature logistique ultime de la croissance scientifique, qui s'exprime par une courbe en S (ou courbe logistique). Voir D. de S. Price, *Science et Suprascience.* Paris : Fayard, 1972 Version originale : *Little Science, Big Science.* Columbia University Press, 1963.

En réalité, le phénomène que l'on peut observer, en fonction de l'analyse quantitative de la littérature scientifique (comptage de revues, d'articles, de résumés, du nombre d'auteurs), est celui des phases de croissance exponentielle qui sont suivies de phases de croissance linéaire.

Le point qu'il nous intéresse de souligner, à propos de la mesure du savoir et de sa croissance par le nombre de publications scientifiques (revues, articles, citations) est qu'il y a, d'une part un problème physique qui concerne directement ceux qui doivent assurer la gestion et le stockage physique des périodiques et des articles, et d'autre part un problème cognitif qui touche directement à l'analyse de l'information.

Selon la «loi de Bradford», nous savons que les sources d'information augmentent selon une progression géométrique, tandis que l'apport d'information au sens cognitif du terme se fait selon une progression arithmétique.

La formule simple de la loi de Bradford proposée par Brookes est :

$$R(n) = k \log (n / s).$$

R(n) est le nombre cumulatif de références ; *n*, le rang du périodique selon sa fréquence ; *k* est une constant qui détermine l'inclinaison de la courbe ; *s* est l'intersection sur l'abscisse (log rang)

Voir à ce sujet S.C. Bradford, "Sources of information on specific subjects" (1934), *Journal of Information Science,* vol. 10, 1985, p. 176-180 ; B.C. Brookes, "Bradford's law and the bibliography of science", *Nature,* vol. 224, 6 December 1969, p. 953-956. M.C Drott et B.C. Griffith, "An empirical examination of Bradford's Law and the scattering of scientific literature", *Journal of the American Society for Information Science,* vol. 29, n° 5, sept. 1978, p. 238-246.

Rappelons que les documents et l'information ne sont pas des entités de même nature.

Lorsque l'on se pose la question de l'analyse de l'information et de sa représentation, nous essayons de rendre visible les structures de connaissance de cette information (dans le monde anglophone on parle de *mapping knowledge structures*), et non pas simplement de compter de documents.

Qu'un lecteur trouve l'information qu'il cherche en consultant un document, c'est la preuve que celle-ci est insérée dans la structure de connaissances du document en question. Bien que le lecteur puisse uniquement extraire les fragments dont il a besoin, il sera tout même rassuré d'avoir trouvé l'information dans un contexte concret où il peut la replacer.

C'est pourquoi il est toujours nécessaire de présenter l'information dans le cadre d'une structure cognitive pertinente. Il est donc important de pouvoir représenter un tel cadre à partir de la connaissance qui est enregistrée dans la littérature scientifique à l'aide d'outils permettant de structurer l'information comme le font les programmes NEURODOC et SDOC.

11. Le réductionnisme bibliométrique.-

L'article scientifique est considéré d'une manière explicite, depuis les années 1960 à peu près, comme un indicateur «output» de la recherche scientifique (le comptage de publications, l'analyse de citations et de co-citations).

Sous la forme "objective" de données bibliographiques, la science devient l'objet empirique d'une approche qui applique l'outil mathématique au "corpus mondial des publications scientifiques", dans lequel se matérialise la connaissance scientifique.

L'article scientifique devient un instrument de définition de la science et du scientifique, une équivalence est ainsi établie entre la notion de science et l'écrit scientifique.

On entend par science ce qui se publie dans les articles des revues, les communications, les rapports, les thèses et les ouvrages scientifiques ; ou d'une manière plus restrictive "la science est ce qui est publié dans les articles scientifiques" (Price, 1969, 94) ; c'est une manière de dire que la science est de la connaissance écrite.

On appelle "scientifique une personne qui a publié un article scientifique" (Price, 1965, 556), "nous définirons un scientifique comme quelqu'un qui quelquefois dans sa vie a aidé à l'écriture d'un article" (Price, 1969, 95). L'idée est que "le produit final majeur du travail d'un scientifique est l'article qu'il publie" (Price, 1969, 94).

Cette réduction, que nous appelons réductionnisme bibliométrique, a permis d'appliquer l'analyse quantitative à l'étude de la science, car la littérature scientifique se prête au dénombrement, à la classification et à la représentation sous la forme de séries temporelles (comme explique Price dans *Little Science, Big Science*).

Le modèle de la science qui sert ici de paradigme est sa représentation comme "une population de publications" où chaque document écrit est considéré "une sorte d'atome de connaissance" (Price, 1969, 92) ; "chaque article représente au moins un quantum d'information scientifique" (Price, 1972, p.70).

Pourtant, et à l'encontre de ce réductionnisme, il faut souligner que "document et connaissance ne sont pas des entités identiques" (comme le rappelle Brookes, 1980, p. 127 : "But document and knowledge are not identical entities).

Bibliographie concernant les citations de Derek J. de Solla Price, *Science et Suprascience.* Paris : Fayard, 1972; "Is Technology Historically Independent of Science ? A Statistical Historiography", *Technology and Culture,* vol. 6 (1965), pp. 553-568; "The Structure of Publication in Science and Technology", in W.H. Gruber et D.G. Marquis (éds.), *Factors in the Transfer of Technology.* Cambridge, Mass. : The MIT Press, 1969, pp. 91-104.

12. Conclusion

Notre but est donc de passer d'un traitement statistique des documents (bibliométrie traditionnelle) à une représentation des connaissances matérialisées dans le langage écrit des données bibliographiques. Actuellement nous utilisons les mots-clés comme une première génération d'indicateurs de connaissance.

En somme, notre intention est le développement d'une «scientométrie qualitative» (selon l'expression de M. Callon, J. Law et A. Rip, voir ch. 7 de *Mapping the Dynamics of Science and Technology*. London, Macmillan, 1986) ou d'une «scientométrie cognitive» (selon l'expression de J-P. Courtial et A. Rip, dans leur article "Co-word Maps of Biotechnology: An Example of Cognitive Scientometrics", *Scientometrics*, vol. 6, 1984, p. 381-400).

Dans la mesure où le but de cette analyse de l'IST est la représentation de la connaissance matérialisée ou objectivée sous la forme de données bibliographiques, des auteurs la considèrent à ce titre comme faisant partie des sciences de la cognition (H.D. White and K.W. McCain, "Bibliometrics", *Annual Review of Information Science and Technology*, vol. 24, 1989, p. 164).

Mars, 1993.

Annexe 2[1]

Une boîte à outils pour le traitement de l'Information Scientifique et Technique.

Cet article présente les idées directrices de la réalisation d'une boite à outils pour le traitement de l'information scientifique et technique (modularité par décomposition en programmes, utilisation du standard SGML), des exemples d'application et des commentaires sur les premiers résultats.

[1] *Ducloy J., Charpentier P., François C., Grivel L. 'Une boîte à outils pour le traitement de l'Information Scientifique et Technique', 4es. Journées Internationales Le Génie logiciel et ses applications. Toulouse, 9-13 Décembre 1991, p. 239-254 ; et dans Génie logiciel, n° 25, p. 80-90, 1991.*

1 - INTRODUCTION

Traditionnellement, la démarche d'informatisation privilégie deux approches. Dans les cas simples, ou plutôt ceux dans lesquels les contraintes transactionnelles sont absentes, on associe simplement un programme à la résolution d'un problème (fig 1). Les mécanismes favorisant la modularité conduisent en fait à une décomposition en fonctions (en utilisant la terminologie Pascal ou langage C), si possible compilées séparément et coordonnées par un programme principal. Les outils et langages associés sont bien connus, même s'ils font encore l'objet de recherches. C'est autour de cette démarche que s'articule la formation à la programmation.

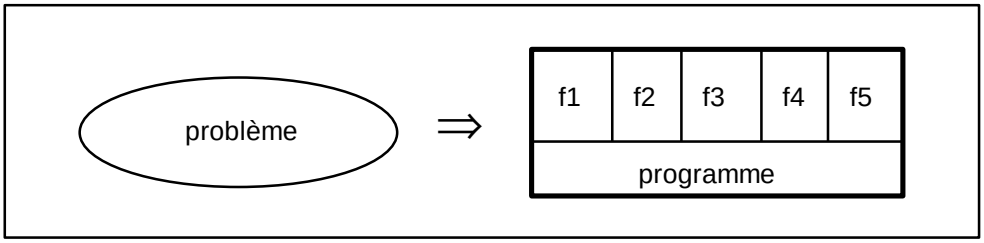


fig 1

Dans les applications présentant des contraintes organisationnelles et transactionnelles, on privilégie une architecture reposant sur une base de données autour de laquelle gravitent commandes transactionnelles (ti) et programmes batch (pi) (fig 2). Les méthodes classiques d'analyse et la formation associée reposent sur cette architecture.

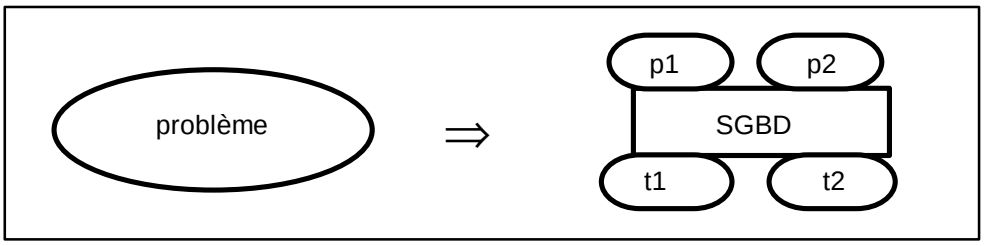


fig 2

L'amélioration de l'ergonomie des systèmes d'exploitation et la simplification des langages de commande permettent de réaliser facilement des communications entre programmes par l'intermédiaire de fichiers, ou par des tubes (ou pipe en terminologie Unix). Pour atteindre l'objectif de modularité, on dispose alors d'un mécanisme complémentaire que nous appellerons "décomposition en programmes". Le problème de la figure 1 qui se décompose en 5 fonctions logiques peut finalement être réalisé en 3 programmes (fig 3), où l'on remarque que la fonction f3 peut être réalisée par un programme spécifique ou par une commande plus générale (telle qu'un tri). Hormis les articles ou ouvrages consacrés à la programmation sous Unix, la littérature, la recherche et la formation sont peu abondantes sur ce sujet.

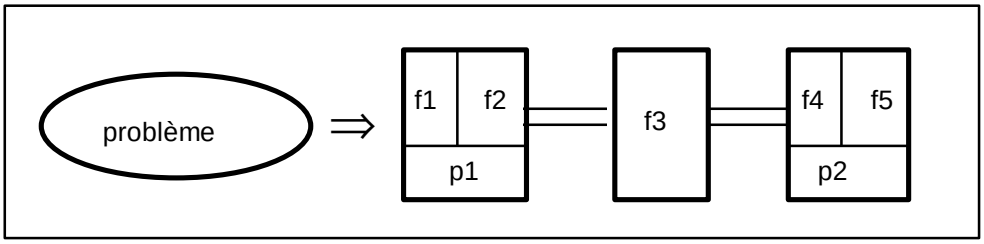


fig 3

Enfin, un problème peut se résoudre en mettant en oeuvre un ensemble de progiciels (par exemple un SGBD et un système documentaire - fig 4). Pour leur permettre de communiquer, on doit souvent réaliser des programmes ou des chaînes de programmes. Ici encore, si ce type d'architecture est de plus en plus répandu, la formation ou les ouvrages méthodologiques ont tendance à l'ignorer.

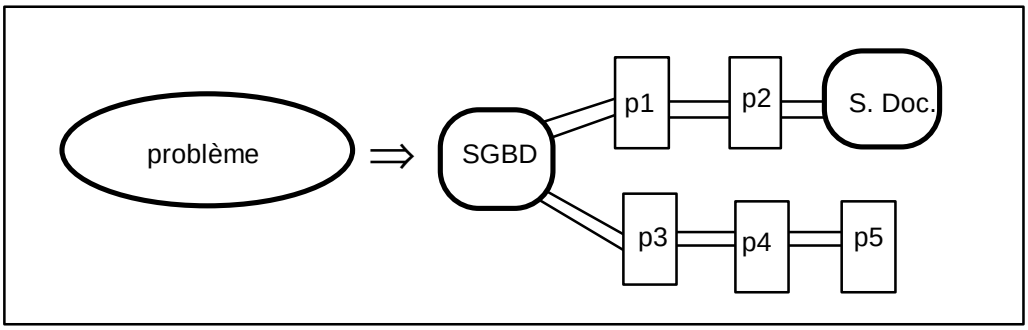


fig 4

L'INIST, centrale documentaire où l'on est amené à traiter tous les problèmes informatiques liés aux sciences de l'information (depuis la bibliothéconomie, jusqu'à l'infométrie en passant par l'édition et la documentation) est perpétuellement confronté à la communication entre progiciels. De plus, les données manipulées "collection de fiches ou notices bibliographiques" se prêtent bien à une modularité par décomposition en programmes qui s'échangent des flux d'information. Le Département Recherche et Produits Nouveaux de l'INIST est en train de réaliser une bibliothèque d'outils d'informatique documentaire, basée principalement sur cette approche.

2 - QUELQUES ASPECTS DE LA MODULARITE PAR DECOMPOSITION EN PROGRAMMES

2.1 - Un exemple d'introduction

Supposons que l'on souhaite analyser un texte de façon à faire apparaître les termes les plus fréquents. Dans une approche classique, ce problème se résout facilement en construisant une liste de couples (terme, fréquence d'apparition). Cette programmation n'a rien de très complexe, mais demande un bon niveau (gestion de listes ou de mémoire, insertion, ...). En utilisant une approche par décomposition en programmes, le problème s'organise alors en 5 étapes (fig 5).

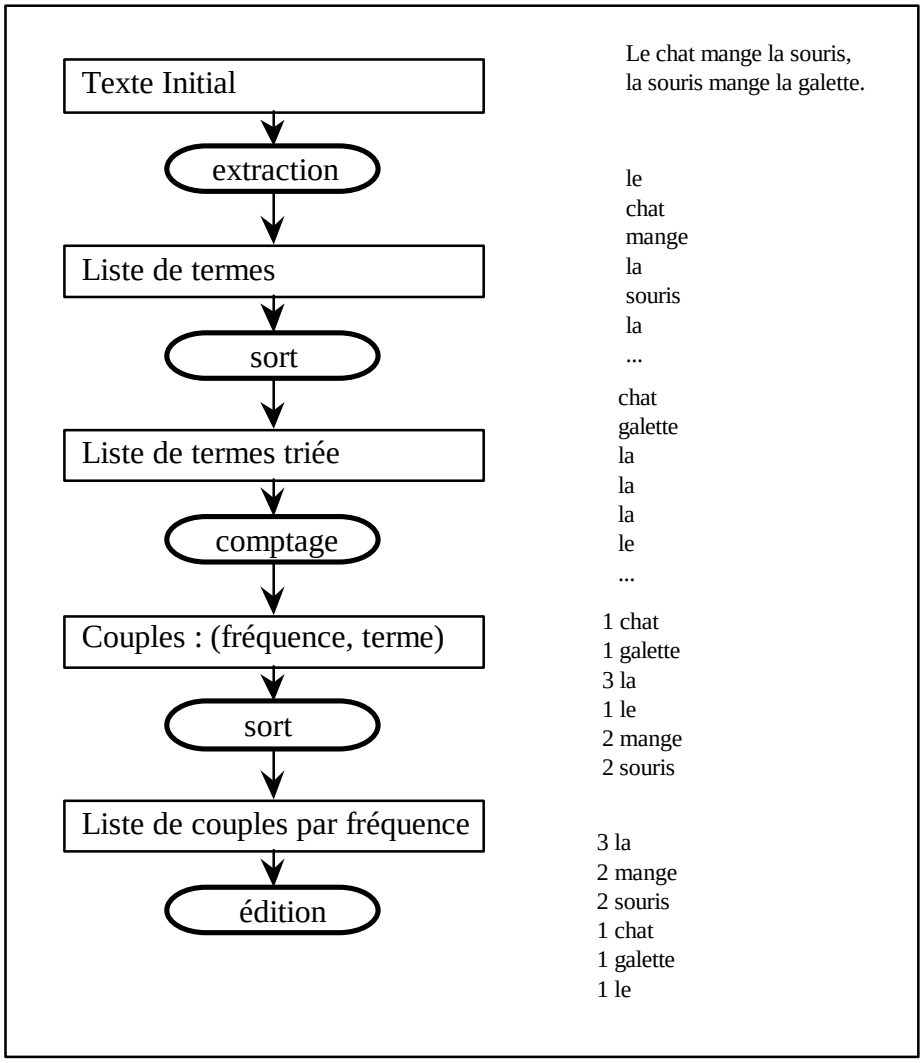


fig 5

Deux étapes utilisent le programme de tri standard, les autres ne présentent aucune difficulté particulière.

2.2 - Quelques qualités de la décomposition en programmes indépendants

Si l'on essaie d'évaluer le résultat obtenu dans le paragraphe précédent, en utilisant par exemple les critères de modularité proposés par B. MEYER[2] [MEY90], on peut vérifier que l'exemple de la figure 5 respecte globalement les 5 critères proposés :

- décomposabilité modulaire : un problème peut être décomposé en sous-problèmes dont la solution peut être recherchée séparément. Les communications entre programmes par d'autres média que les fichiers sont très peu commodes, et cette technique oblige donc à découper un problème en sous-problèmes qui se réduisent à passer d'un fichier x à un fichier y.

[2] en toute rigueur, ces critères s'appliquent à une méthode et non à un outil.

- composabilité modulaire : la méthode favorise la production d'éléments qui peuvent être combinés les uns avec les autres pour produire de nouveaux systèmes. Les modules "comptage" ou "extraction" de l'exemple précédent peuvent être réutilisés dans n'importe quelle application.

- compréhensibilité modulaire : la méthode aide à produire des modules dont chacun peut être compris séparément par un lecteur humain.

- continuité modulaire : une petite modification de la spécification du système n'amène à modifier qu'un seul module [...]

- protection modulaire : l'effet d'une condition anormale, se produisant à l'exécution, reste localisé à ce module. Une condition anormale ne peut se propager d'un programme au suivant que si, n'ayant pas été détectée, elle induit une séquence non conforme aux spécifications dans le flot de données. Hormis ce cas, le critère de protection modulaire est respecté, en particulier tous les incidents dûs aux effets de bord, allocation de mémoire ou débordements de tableaux sont purement localisés à un programme.

2.3 - Les conditions d'application de la décomposition en programmes indépendants

2.3.1 - Problèmes (ou sous-problèmes) avec peu de contraintes transactionnelles

Cette technique de décomposition était bien connue en informatique de gestion dans les années 60, elle a montré ses limites dès qu'une application devenait complexe (contraintes transactionnelles, nombreux aspects organisationnels). Mais les techniques n'utilisant qu'une seule approche SGBD - L4G ont également leurs limites (par exemple, traitement des informations de taille variable - données textuelles -, problèmes à forte contraintes algorithmiques, prototypage, arbres à profondeur variable, etc). Les avancées technologiques ont fait reculer certaines barrières ; par exemple sur une station de travail, il faut moins d'une seconde pour soumettre et exécuter une série de commandes sur un ensemble de fichiers. Dans l'édition, la bibliothéconomie ou la documentation, les délais de fabrication ou d'élaboration intellectuelle des informations sont parfois de plusieurs jours, et quelques minutes de délai dans le traitement de certaines transactions n'ont aucune autre incidence.

On peut donc parfaitement mettre en place des architectures mixtes, où une partie des informations est gérée par un SGBD et les traitements à caractère algorithmique sont effectués par des chaînes de commandes.

2.3.2 - Système proposant des mécanismes simples de communication entre programmes (Unix ou équivalent)

Il serait absurde de chercher à décomposer un problème en modules de quelques lignes si l'assemblage des modules devenait plus complexe que le corps des modules eux-mêmes. Les systèmes d'exploitation anciens (MVS par exemple) s'avèrent donc inadaptés.

Parmi les systèmes actuels, une implémentation correcte du mécanisme de pipe (communication par buffers et non simulée par fichiers) s'avère rapidement indispensable, pour éviter une trop forte expansion du volume de données de départ (dans l'exemple précédent, en dehors des fichiers de travail du tri, on atteint un facteur 4) ou une multiplication comparable du nombre d'accès disque (même type de rapport).

2.3.3 - Utilisation d'outils lexicaux et syntaxiques

La décomposition en programmes conduit pour chaque commande à traiter l'analyse des entrées et le formatage des sorties. Ici encore, tout le bénéfice de la décomposition peut être perdu si l'on n'utilise pas de techniques adaptées. Les générateurs d'analyseur lexical Lex et d'analyseur syntaxique Yacc, même s'ils ont été conçus pour d'autres objectifs, s'avèrent particulièrement efficaces pour l'analyse des entrées.

Par exemple, si les textes sont simples, et en considérant que l'on ne fait pas de traitement lexicographique complexe, le programme d'extraction des termes du texte initial de la figure 5, se réduit à deux règles Lex :

```
%%
[ ,;.:\n\t]+        printf("\n");
                    /* remplacer toute chaîne de séparateurs par un saut de
ligne */
[A-Z]               printf("%c", tolower(yytext));
                    /* conversion des majuscules en minuscules */
%%
```

Remarquons la simplicité de ce module d'extraction qui se résume à de simples règles de transformation de caractères. Le regroupement des termes en vue du comptage est effectué par la commande de tri (sort sous Unix).

2.3.4 - Utilisation de balisages parenthésés et descriptifs (SGML ou équivalent)

Pour profiter au mieux des avantages offerts par les outils lexicaux, et dans la mesure où le programmeur possède la totale maîtrise des spécifications des données échangées entre programmes, il paraît opportun d'examiner le formatage de ces données.

En fait, seule l'analyse des données pose un problème de reconnaissance. La partie dédiée à l'analyse du flot de données peut devenir rapidement un "programme spaghetti" non maintenable pour peu que plusieurs tests doivent être réalisés simultanément. Par exemple : se demander si le caractère lu est un caractère courant de la zone en cours, le caractère de fin, le caractère de fin d'une zone englobante, le caractère de début de la zone suivante (et dans ce cas le caractère précédent était, mais on ne le sait que maintenant, le caractère de fin de la zone précédente) !

Pour éviter cet inconvénient majeur, il suffit d'être rigoureux dans les spécifications des données intermédiaires. Cette spécification peut être souvent décrite à l'aide d'une

grammaire. La norme SGML [ISO 8879-1986] nous offre un outil pour la décrire : la DTD (Document Type Definition), [HER88].

SGML (Standard Generalized Markup Language), conçu comme un format d'échange pour documents électroniques en vue de leur impression, reprise dans le projet CALS[3] du DOD (Department of Defense, USA) , est en réalité d'un usage beaucoup plus général, [BOR90], [EWG90], [NEW90]. En effet, SGML donne des règles de balisage pour décrire des structures arborescentes où chaque noeud est identifié par une étiquette. Baliser un document consiste à insérer dans le texte des chaînes de caractères qui donnent de l'information sur le contenu du document.

A titre d'exemple, une notice bibliographique provenant d'un serveur ou d'un CD/ROM se présente généralement comme suit :

NO : 90-0128293
TI : Density-dependent interactions between seedlings of Dactylorhiza majalis (Orchidaceae) in symbiotic in vitro culture
AU : RASMUSSEN (H.);JOHANSEN (B.);ANDERSEN (T. F.)
...

La structure logique d'une telle information est très simple (une suite de champs repérés par un identifieur) et en suivant la norme SGML, on peut lui associer une DTD élémentaire telle que :

```
<!ELEMENT record ... (NO, TI, AU, ...)>
```

Il est relativement facile de définir les règles lexicales qui permettent d'identifier le début ou la fin d'une notice, le début ou la fin d'un champ à l'intérieur de la notice de manière à la transformer en document SGML en forme normale.

En forme normale SGML, chaque champ est repéré par une balise de début : <generic_identifier> et une balise de fin : </generic_identifier>. En utilisant un "parser normalizer", ou en écrivant un programme Lex (générateur d'analyseur lexical), la notice ci-dessus est transformée en format SGML comme suit :

```
<record>
<NO>90-0128293</NO>
<TI>Density-dependent interactions between seedlings of Dactylorhiza
majalis (Orchidaceae) in symbiotic in vitro culture</TI>
<AU>RASMUSSEN (H.);JOHANSEN (B.);ANDERSEN (T. F.)</AU>
...
</record>
```

Cette forme obtenue, la plupart des traitements sur de tels documents se réduisent à associer des actions à un élément de la grammaire et, dans bien des cas, travailler au

[3] CALS : Computer-aided Acquisition and Logistics Support.

niveau lexicographique suffit. Par exemple, la phase "extraction" des termes de l'exemple de la figure 5 se réécrit comme suit :

```
%START SAUT, ECLATEMENT
%%
"<TI>"                          BEGIN ECLATEMENT ;
"</TI>"                         BEGIN SAUT ;
<ECLATEMENT>[ ,;.:\n\t]+            printf("\n");
<ECLATEMENT>[A-Z]                       printf("%c",
tolower(yytext));
<SAUT>                              ;
%%
main()
{BEGIN SAUT; yylex();}
```

L'utilisation de Lex permet d'associer facilement des actions (IMPRESSION) ou des états (SAUT et ECLATEMENT) lorsque l'on rencontre une balise.

L'usage simultané de systèmes d'exploitation intégrant correctement le pipe, l'utilisation d'outils lexicaux ou syntaxiques sur des structures balisées offre donc une base technique à une décomposition modulaire basée sur la communication par tube. Pour aller plus loin nous développons actuellement une bibliothèque basée sur ce concept.

3 - ILIB, UNE BIBLIOTHEQUE DE MODULES ET DE FONCTIONS AUTOUR DE LA NORME SGML

3.1 - Le domaine d'application de la ILIB

Le domaine d'application prioritaire de cette bibliothèque est la fabrication d'informations élaborées à partir des bases de données documentaires ou factuelles d'origine diverse, internes à l'INIST (bases PASCAL, FRANCIS, WTI) ou extérieures.

Ces informations élaborées de nature très diverse (depuis de simples documents papier jusqu'aux hypertextes) sont obtenues par des traitements linguistiques et statistiques sur des sous-ensembles de documents extraits de ces bases.

Une des premières difficultés de ce type d'application provient de la multitude de formats et de structures de données qu'il faut manipuler, analyser, croiser ou éditer. En revanche, on peut constater que ce type d'application n'a pratiquement aucune contrainte transactionnelle. C'est donc un domaine privilégié de la décomposition en programmes indépendants.

3.2 - Modèle de données et utilisation de la norme SGML

Pour faciliter les spécifications des éléments de la bibliothèque il s'est avéré intéressant d'utiliser un modèle de données en couche. Les couches doivent être indépendantes les unes des autres, et un outil de la bibliothèque ne doit travailler que sur une couche à la fois.

De la plus basse à la plus haute, les différentes couches sont décrites dans la figure 6.

4	base de données
3	enregistrement
2	éléments de données
1	objets élémentaires

fig 6

3.2.1 - Niveau des objets élémentaires

Cette couche ne concerne que les règles de codification des objets élémentaires d'un point de vue matériel (caractères, entiers, ...).

Comme il s'agit d'échanges de données entre processus pouvant s'exécuter sur des systèmes différents, nous nous sommes limités au seul type caractère. En pratique nous avons défini un jeu de caractères minimal, ne posant aucun problème de visualisation sur imprimante ou terminaux (sous-ensemble des caractères graphiques de la norme ISO 646), ayant comme seuls caractères de contrôle le saut de ligne (séparateur d'enregistrements) et la tabulation (séparateur de zones).

3.2.2 - Niveau élément de données

Cette couche est utilisée pour spécifier les objets élémentaires au niveau d'une application (codification des nombres, des dates, des noms de pays, des noms propres, ...). Pour le moment, nous nous sommes surtout intéressés à la codification des caractères spéciaux et accentués utilisés dans les langues latines.

L'annexe D.4 de la norme SGML propose une suite de recommandations pour coder les caractères diacritiques, les caractères accentués des langues latines, grecques ou cyrilliques. Chaque caractère est représenté par un "et commercial" (&) suivi par son identification et un point virgule. Par exemple é est représenté par é, â par â, a par &agr;. Tous les modules linguistiques, de préparation de tris ou d'édition de la bibliothèque utilisent cette recommandation et la phrase suivante :

> Les normes Unimarc & SGML sont utilisées pour l'échange de données, ¹ = 3.14159.

peut être codée comme suit :

> Les normes Unimarc & SGML sont utilisées pour l'échange de données, &pgr; = 3.14159.

3.2.3 - Niveau notice ou enregistrement

A ce niveau sont traitées les structures composées d'éléments simples mais manipulées de façon globale au niveau des entrées-sorties. C'est à ce niveau que l'on trouve par exemple la description des formats de notices bibliographiques. Il s'agit donc simplement de définir pour chaque type de données une DTD SGML.

Les formats d'échanges des informations bibliographiques (Unimarc, Pascal, CCF, ...) préconisent des structures à 2 niveaux suivant la norme ISO 2709. Certaines organisations (par exemple la CEE avec FORMEX [EC85] [GUI90]) proposent des DTD qui reprennent toutes les informations bibliographiques d'un format particulier, mais dans une organisation spécifique. Nous avons choisi une approche différente en définissant une DTD directement associée au format ISO 2709 [FRA90], permettant d'écrire un programme de transformation s'appliquant de fait à tous les formats dérivés [DUS91] (fig 7).

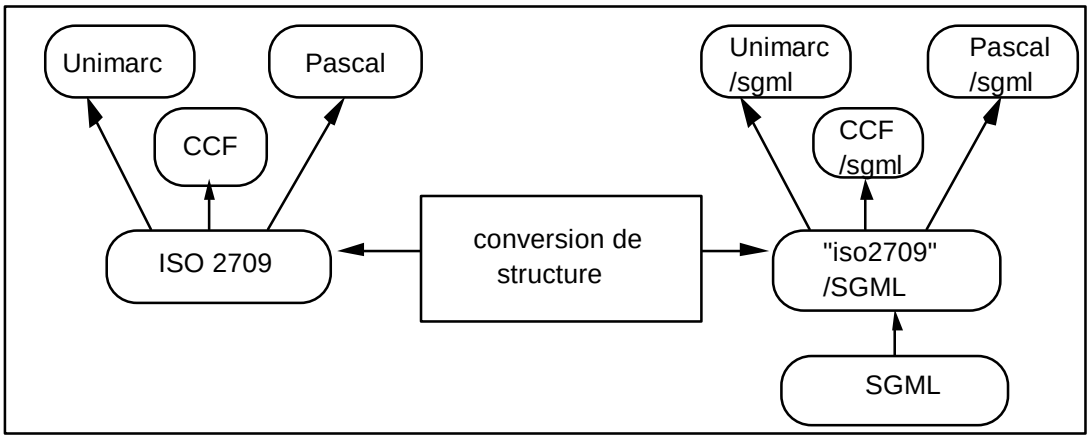


fig 7

Par exemple l'information "collectivité auteur" d'une notice Unimarc, classiquement éditée en forme externe comme suit :

210 $aLisboa$cMuseu Nacional de Arqueologia e Etnologia$d1895

produit la structure SGML suivante :

```
<f210>
        <sa>Lisboa</sa>
        <sc>Museu Nacional de Arqueologia e Etnologia</sc>
        <sd>1895</sd>
</f210>
```

De façon similaire on peut facilement définir une structure SGML équivalente aux formats obtenus par déchargement de serveurs ou de CD/ROM. Par exemple la notice suivante :

NO : 90-0128293
TI : Density-dependent interactions between seedlings of Dactylorhiza majalis (Orchidaceae) in symbiotic in vitro culture
AU : RASMUSSEN (H.);JOHANSEN (B.);ANDERSEN (T. F.)

AF : N1 Univ. Copenhagen, botanical lab.;V1 Copenhagen 1123;P1 DNK
DT : Publication en serie

devient en SGML :

```
<record>
<NO>90-0128293</NO>
<TI>Density-dependent interactions between seedlings of Dactylorhiza
majalis (Orchidaceae) in symbiotic in vitro culture</TI>
<AU>RASMUSSEN (H.);JOHANSEN (B.);ANDERSEN (T. F.)</AU>
<AF>
    <N1>Univ. Copenhagen, botanical lab.</N1>
    <V1>Copenhagen 1123</V1>
    <P1>DNK</P1>
</AF>
<DT>Publication en serie</DT>
</record>
```

De même les données intermédiaires sont spécifiées en utilisant la norme SGML. Voici un exemple de liste inverse :

```
<idx>
        <kw>computer</kw>
        <f>3</f>
        <l>001245,015254,025487</l>
</idx>
```

où kw est le libellé du mot-clé, f la fréquence d'indexation et l la liste de références de documents indexés.

3.2.4 - Niveau base ou ensemble de données

A ce niveau sont définis les ensembles de données de type divers rencontrés dans une application (bases de données, bases documentaires, fichiers séquentiels, hiérarchisés, ...). Par exemple, nous avons défini une organisation des fichiers par accès direct, facilement manipulables par l'utilisateur et par programme. Les enregistrements y sont regroupés en fichiers et répertoires de 100 éléments de façon hiérarchique. Par exemple, l'enregistrement 014825 est le 26ème enregistrement du fichier 48.file qui se trouve dans le répertoire 01.dir.

Ce type d'organisation convient en particulier aux applications non transactionnelles dans lesquelles il n'y a pas de mise à jour.

3.3 - Principales fonctions de la bibliothèque

La structure d'un fichier SGML ayant de nombreux points communs avec un programme structuré, la boîte à outils standard d'Unix s'applique donc à SGML. Ainsi beaucoup de fonctions sont inspirées de ces outils. D'autre part, les générateurs d'analyseurs lexicaux ou syntaxiques (Lex et Yacc) se comportent comme de

véritables outils de génie logiciel pour composer des programmes travaillant sur des données SGML.

3.3.1 - Des modules de conversion

Une des premières étapes de toute application est de convertir les documents de provenance quelconque dans le formalisme décrit dans le modèle ci-dessus. Il existe donc un ensemble de fonctions de conversion des divers formats vers SGML.

On trouve des modules de conversion adaptés aux trois couches les plus basses du modèle représenté sur la figure 6. Dans la première couche sont simplement traitées les conversions de types ascii <-> ebcdic, dans le deuxième on trouve une collection de conversions de jeux de caractères (latins, grecs cyrilliques, ...) vers SGML, enfin dans la troisième la structure des enregistrements est convertie.

3.3.2 - La construction d'ensembles documentaires

Lorsque les données sont uniformisées, nous pouvons créer des fichiers directs avec l'organisation décrite dans le paragraphe 3.2.4 ce qui nous permet d'avoir un accès direct aux enregistrements.

Ensuite nous pouvons construire des fichiers inverses à partir d'un champ choisi dans le fichier direct (mot-clé, auteur, ...). Il est possible d'appliquer des filtres linguistiques pour une indexation automatique.

Les chaînes qui permettent de construire un tel ensemble documentaire à partir de références quelconques utilisent en fait des modules de la bibliothèque et des commandes de base. Il est très facile d'y insérer un filtre spécifique écrit en Lex.

3.3.3 - Des modules applicatifs

A partir des fichiers directs et des fichiers inverses, des études bibliométriques ou scientométriques peuvent être menées et des applications telles que celles présentées dans le chapitre 4 peuvent être développées. Pour cela des fonctions d'accès aux données adaptées aux documents structurés en SGML se sont avérées nécessaires.

Par exemple, une fonction, largement inspirée de la philosophie de la commande grep d'Unix, permet de créer un nouveau fichier en sélectionnant ou en éliminant des enregistrements qui contiennent une certaine forme (ou pattern). Ainsi, utiliser cette fonction peut servir à éliminer, dans un fichier inverse, les enregistrements correspondant à une fréquence inférieure à un certain seuil.

3.4 - Intégration à la philosophie Unix

3.4.1 - La paramétrisation des fonctions

Quel que soit le type d'information initiale (Unimarc, CCF, format élémentaire, …), le mécanisme de structuration est unique en SGML et cela quel que soit le niveau d'un élément dans l'arborescence. Il est donc possible de définir un opérateur capable de

faire des manipulations sur des arbres ou des éléments d'arbre quel que soit la localisation d'un élément dans cet arbre. De cette constatation est née l'idée de paramétrer certaines fonctions avec des options standardisées.

Par exemple, tous les filtres qui opèrent sur un élément spécifique d'une structure SGML utilisent l'option -m (pour mark) associée à un identificateur de balise ou à un chemin de balises (suite de balises séparées par des caractères "/" par analogie avec le mécanisme d'adressage d'Unix). De plus, le motif décrivant ce chemin peut être exprimé à l'aide de métacaractères.

Par exemple :

* signifie zéro ou n occurrences de n'importe quel caractère
? signifie une occurrence de n'importe quel caractère
les crochets [et] permettent d'exprimer une liste de caractères ; [a-z] désigne donc une lettre de l'alphabet en minuscule et [0-9] un chiffre

Ainsi, la séquence inspirée de la commande grep d'Unix : Sgmlgrep -e apple -m kw sur le fichier :

```
<record><kw>apple,orange,lemon</kw><title>fruit</title></record>
<record><kw>plane tree,poplar,oak,beech</kw><title>tree</title></record>
```

permet de sélectionner les enregistrements comportant le mot "apple" sous la balise kw :

```
<record><kw>apple,orange,lemon</kw><title>fruit</title></record>
```

En pratique, il existe dans la bibliothèque un ensemble de fonctions qui analysent l'effet d'une option m sur un fichier SGML.

On remarquera que la plupart de ces outils qui sont souvent des opérateurs élémentaires n'utilisent qu'un niveau lexicographique et ne demandent donc pas un paramétrage par une grammaire complète. Autrement dit, un générateur comme Lex suffit ; un parser SGML ou un générateur d'analyseur syntaxique comme Yacc sont inutiles (ou même inutilisables) à ce niveau.
En revanche, ils sont utiles voire parfois indispensables pour convertir des documents complexes en structures normalisées (balisage maximum), ou pour écrire des traitements spécifiques s'appliquant à un type précis de documents.

3.4.2 - La documentation

En plus de la documentation du style "manuel utilisateur", nous nous sommes inspirés d'Unix pour faire une documentation pour chaque fonction dont la forme est typiquement celle des "man" sous Unix.

4 - EXEMPLES D'APPLICATIONS DE LA BIBLIOTHEQUE

4.1 - Infométrie et hypertextes

L'infométrie est un terme utilisé pour couvrir les techniques utilisées pour maîtriser la complexité d'ensembles de données en mettant en évidence des concepts ou des thèmes dominants. Plus précisément, on peut citer la bibliométrie qui sert à évaluer des fonds bibliographiques, la scientométrie qui a pour vocation de fournir des indicateurs pour l'évaluation de la R&D, les outils d'aide à la veille scientifique. Les études infométriques sont essentiellement basées sur des analyses statistiques ou plus précisément d'analyse de données (classification).

L'hypertexte a pour vocation d'articuler et d'organiser des composants élémentaires d'information sous forme de réseaux de connaissance, à l'aide de noeuds contenant de l'information multimédia (texte, image, graphique, son) interconnectés par des liens représentant les relations existantes entre ces granules de connaissance [DAN90].

La combinaison de ces deux techniques (calcul de graphe par infométrie et visualisation par hypertextes) permet à un utilisateur de naviguer à travers l'information pertinente en suivant les relations entre concepts établies d'un point de vue statistique.

Le Centre de Sociologie de l'Innovation de l'Ecole des Mines de Paris et le CDST[4] ont défini et mis au point plusieurs approches d'analyses de données pour des études scientométriques. Elles ont été expérimentées avec succès sur Macintosh et sur PC, pour traiter des volumes de données moyens (20 à 30 000 documents) [MIC88]. Elle ont été redéfinies et réécrites dans le cadre de la ILIB, donnant naissance à deux applications, SDOC[5] et NEURODOC[6]. .

4.2 - Le projet SDOC : les cartes conceptuelles

L'application SDOC [GRI91] produit des cartes montrant l'organisation conceptuelle d'un domaine scientifique à partir d'un sous ensemble de notices bibliographiques extraites de bases de données telles que PASCAL ou FRANCIS. SDOC repose sur la méthode des mots associés - analyse des cooccurrences des mots-clés - qui a déjà été exploitée dans le cadre du produit LEXIMAPPE[7] .

Un indice statistique permet de mesurer la force associative de deux mots-clés. Cet indice est une fonction du nombre d'occurrences du chacun des termes, et du nombre de cooccurrences des deux termes. L'ensemble des associations entre mots-clés forme un réseau valué d'associations. Un algorithme de classification basé sur la

---

[4] ancien centre de documentation du CNRS avant la création de l'INIST.

[5] SDOC bénéficie d'un financement de la CEE (projet ESPRIT KWICK n° 2466).

[6] Ce projet bénéficie d'un financement du MRT et du SERICS dans le cadre de l'appel à propositions "Interfaces Intelligentes".

[7] LEXIMAPPE est une marque déposée du CNRS et de l'Ecole des Mines de Paris.

méthode du simple lien ([CAL83] et [MIC88]) permet de découper le réseau en clusters (groupement de mots avec des relations entre ces mots). D'un point de vue sémantique, les études réalisées montrent que les clusters s'apparentent aux thèmes de recherche que l'on peut trouver dans un domaine scientifique. De plus, les clusters peuvent admettre des relations avec d'autres clusters et à chaque cluster est associée une liste (triée par degré de pertinence) de références bibliographiques.

On obtient donc un réseau structuré et hiérarchisé de clusters (par opposition au réseau "plat" des associations entre mots). Ce réseau de clusters est ensuite traduit en termes de noeuds et de liens hypertextes. Cette représentation permet à un utilisateur de naviguer de thèmes en thèmes, "d'îlots de connaissance en îlots de connaissance", puis via les thèmes, d'accéder aux références bibliographiques.

4.3 - Le projet "NEURODOC : nouveaux profils documentaires"

Le logiciel "NEURODOC" [LEL90] effectue une classification simultanée des documents et des mots-clés qui les indexent. Il extrait de la base de références bibliographiques un ensemble de thèmes. Chacun d'entre eux est représenté par un axe sur lequel se regroupent et s'ordonnent à la fois les documents et les mots-clés.

Ces thèmes correspondent à un type de classe particulier :

- ces classes sont recouvrantes, car un document ou un mot-clé peut appartenir à plusieurs classes à la fois ;
- les éléments, documents et mots-clés, de chaque classe sont ordonnés selon un degré de ressemblance au type idéal de la classe.

Les thèmes sont situés les uns par rapport aux autres sur une carte globale présentée à l'utilisateur dans le dispositif d'interface hypertexte utilisé. Cette carte globale des thèmes est réalisée par une projection sur un plan, des thèmes représentés dans l'espace des mots-clés.

4.4 - Utilisation pratique des résultats infométriques

Les méthodes sous-jacentes à SDOC et NEURODOC produisent des résultats de même type, mais qui peuvent présenter des différences notables tant dans les concepts mis en évidence, les regroupements entre documents ou leurs relations. En pratique, nous avons pour l'instant deux cibles privilégiées.

D'une part, SDOC et NEURODOC peuvent être utilisés pour de l'investigation en Information Scientifique et Technique : observation ou évaluation de fonds documentaires, recherche d'émergence de concepts en veille scientifique, construction a posteriori de thésaurus. Dans ces exemples, il s'agit généralement de prendre un ensemble de références bibliographiques ou de brevets et de leur appliquer une série d'analyses [DUC 91-1].

D'autre part, nous utilisons SDOC et NEURODOC pour construire de nouveaux produits documentaires. Actuellement, l'INIST fournit à ses clients des profils documentaires qui correspondent aux listes des références bibliographiques les plus récentes concernant le domaine choisi. Nous nous proposons de fournir à l'utilisateur

un document hypertexte dans lequel l'ensemble de ces références est complété par un outil de navigation dont le noyau est une carte globale des thèmes [DUC91-2]. En pratique, la cible prioritaire est la création de piles Hypercard pour Macintosh.

La constitution d'un hyperdocument peut être découpée en quatre étapes (fig 8) :

- extraction des documents,
- reformatage des documents,
- traitements statistiques,
- enfin, édition (affichage, mise à jour et impression).

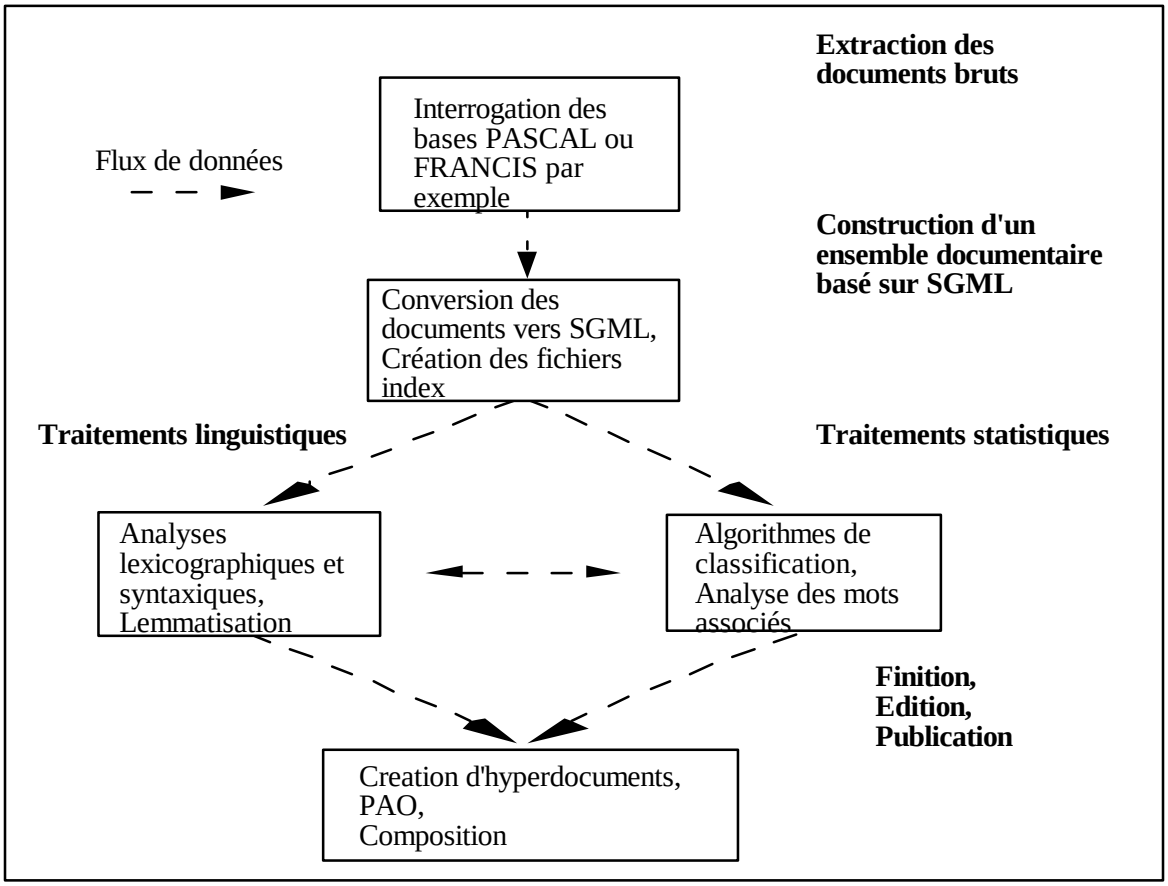


fig 8

Les deux applications décrites ci-dessus diffèrent essentiellement par les méthodes statistiques utilisées et les interfaces hypertextes choisies. A chaque méthode correspond un module de la ILIB (en réalité le même principe de décomposition a été utilisé). La communication entre les programmes de cette architecture est basée sur la définition précise de la structure des données intermédiaires. Les deux premières étapes de la figure 8 s'appuient sur le modèle de données (fichiers inverses et fichiers d'associations). De la même façon, nous avons défini une structure SGML commune décrivant les "clusters" ou "classes" obtenus par les modules statistiques des applications SDOC et NEURODOC.

Une fois les clusters obtenus, il faut les éditer et/ou les publier. Là encore, le même mécanisme s'applique et il est extrêmement facile d'associer à une balise un traitement pour éditer son contenu ou pour créer un nœud hypertexte. En outre, cette facilité d'écriture garantit la possibilité pour les deux applications d'échanger facilement les

“cibles” (hypertexte sous Unix, sous Macintosh, sous PC, publication sous NROFF, Latex, ...). De même, la formalisation de la structure des données intermédiaires permet de remplacer un module de traitement statistique par un autre ou d'intégrer facilement une étape de traitement linguistique dans le processus de fabrication d'informations élaborées.

5 - QUELQUES PREMIERES OBSERVATIONS

Une première version de cette bibliothèque a été mise en service interne au Département au premier trimestre 91, elle a commencé à être utilisée pour des applications réelles en fin de premier semestre 91. Notre expérience est donc réduite mais des premières observations peuvent déjà être dégagées.

Les résultats escomptés en matière de réutilisation ont été atteints, c'est-à-dire qu'une fonction mise en bibliothèque est effectivement utilisée par d'autres. Cependant, ce résultat est loin d'être gratuit. Nous observons très souvent un facteur multiplicatif supérieur à 10 entre l'écriture d'un programme permettant de résoudre un problème précis et l'obtention d'un module de bibliothèque correctement documenté. De plus, l'écriture de fonctions de bibliothèque demande des informaticiens très confirmés. Plus précisément nous obtenons les ordres de grandeur suivants :

- écriture d'un programme par un programmeur : 3 jours,
- écriture du même programme par un informaticien de haut niveau : 1/2 journée,
- conception, écriture et documentation d'un module de bibliothèque par ce spécialiste : 1 semaine.

La décomposition en programmes s'avère extrêmement performante lors de la phase de mise au point des applications. D'une part, parce que les programmes sont petits (si l'on utilise effectivement des outils lexicaux). D'autre part, parce que, en cas d'incident, les sorties intermédiaires peuvent être redirigées sur un fichier, où il est facile de repérer précisément l'incident, de l'isoler et de le reproduire.

Cette décomposition permet également de faire reculer des contraintes physiques. Par exemple, dans la version antérieure de l'application SDOC, la matrice des cooccurrences des paires de mots-clés était construite en mémoire centrale, ce qui limitait le nombre de documents que l’application pouvait traiter. En combinant de simples programmes Lex indépendants avec des tris (comme dans la figure 5) on obtient le fichier de cooccurrences des mots-clés directement à partir du flot de données, sans rien stocker en mémoire centrale. Cette amélioration est importante étant donné le volume de données à traiter. En effet, les bases PASCAL et FRANCIS contiennent des millions de références, et un domaine peut concerner des centaines de milliers de références bibliographiques.

Enfin, une dernière observation, plutôt inattendue au départ. L'utilisation d'un balisage descriptif s'est avérée très performante dans le dialogue avec les utilisateurs qui peuvent très rapidement lire et interpréter une structure SGML. Nous avons pu mener des opérations d'analyse de données en vue de veille scientifique en travaillant uniquement sur des données (notices, listes inverses, associations ou clusters) en balisage SGML brut. De même, du côté des informaticiens et concepteurs nous avons

pu constater que la manipulation ou la visualisation de données intermédiaires en format SGML était un support plus performant pour l'intuition que la simple spécification abstraite de ces mêmes informations.

6 - VERS DES ATELIERS

En théorie, cette bibliothèque est un maillon d'un ensemble plus complexe illustré par la figure ci-dessous. Les notices bibliographiques sont élaborées dans le cadre d'un schéma relativement classique (production des notices sur station de travail [COR91] et gestion de la production sur SGBD [DUC89]), puis recopiées pour être exploitées sur une plate-forme documentaire.

Au niveau de la production, un atelier de génie logiciel classique est parfaitement adapté. Au niveau de l'exploitation, la bibliothèque ILIB constitue le noyau de ce qui devrait devenir un Atelier Flexible pour la Fabrication d'Informations Elaborées. Enfin, pour bâtir cet Atelier, en utilisant les résultats précédents peut-être faut-il construire ou plus précisément adapter un Atelier de Génie Logiciel. La construction de ces deux ateliers est l'objet de ce paragraphe (il est à noter que l'INIST a pour mission de produire de l'information et non du logiciel, et que des actions de développement en Génie Logiciel ne sont entreprises que si cela s'avère strictement nécessaire).

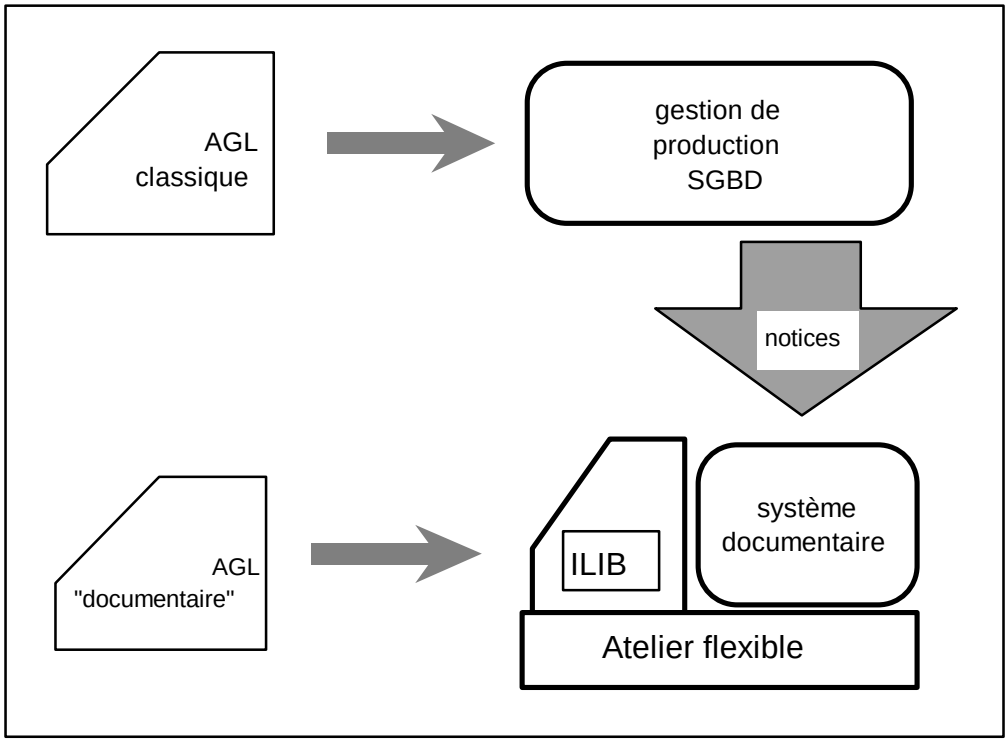


fig 9

6.1 - Atelier Flexible pour la fabrication d'Informations Elaborées

L'Atelier Flexible doit permettre à un ingénieur en sciences de l'information de réaliser rapidement (“à la demande” !) une simple investigation, une étude ou une chaîne documentaire complète capable de produire des documents bruts (notices

bibliographiques, ...) comme des documents plus élaborés (cartes conceptuelles, synthèses bibliographiques, ...). La bibliothèque est la première pierre de cet objectif. Trois types d'extensions sont en cours.

D'abord, on étend ses fonctionnalités par de nombreux outils (nouveaux modules statistiques, linguistiques) et de nouvelles techniques (ou des passerelles vers des produits du marché) telles que OCR, traitements graphiques, ... Ensuite, nous complétons la bibliothèque strictement logicielle par une partie "documentaire", en particulier par une collection d'échantillons de notices de provenances diverses accompagnées de leurs modules de conversion en SGML et d'exemples de traitements.

Enfin, nous abordons l'exploitation d'ensembles de données volumineux. La mise en œuvre de techniques simples telles que celles citées au paragraphe 3.2.4 donne des résultats très prometteurs. Le traitement de 1000 notices Unimarc (environ 2000 caractères) demande une minute sur station SUN (y compris la transformation en SGML et la création de fichiers inverses) ; il faut par exemple moins d'une heure pour obtenir une version à jour du catalogue des périodiques à partir d'une bande magnétique, alors que cette opération demande plusieurs jours sur la machine de gestion en amont ; ceci pour un coût de stockage minime. Pour l'exploitation des données, (et uniquement pour l'exploitation), nous sommes donc amenés à proposer, à côté d'une architecture SGBD, un ensemble de bases d'exploitation éventuellement redondantes (c'est-à-dire qu'un même ensemble d'informations peut exister plusieurs fois avec des structurations différentes). Nous travaillons en ce moment sur ce type d'organisation, où les outils traditionnels tels que dictionnaires de données doivent être adaptés pour tenir compte des nombreuses redondances volontairement acceptées ou pour intégrer une description parallèle dans les DTD SGML.

Pour obtenir un réel Atelier Flexible, il faudrait enfin prendre en compte l'interface homme-machine, et nous serons probablement amenés à avoir deux stratégies. Pour les opérations bien maîtrisées et dédiées à des utilisateurs ciblés (fabrication répétitive mais paramétrable d'un produit), nous produirons, en utilisant des solutions existantes (Aida ...) des ateliers "peu flexibles" mais utilisables par des non-informaticiens. Pour les opérations complexes, véritables investigations documentaires, l'expérience montre qu'il y a toujours des petits développements informatiques à réaliser, et donc qu'elles doivent être menées par du personnel ayant une forte compétence informatique, et les environnements courants, tels que SunView, même s'ils méritent des améliorations s'avèrent parfaitement adaptés.

6.2 - Quels Ateliers de Génie Logiciel pour l'ingénierie de l'IST ?

Nous avons choisi délibérément d'investir fortement sur le niveau constituants de base, par exemple fichier mots-clés et fichier cooccurrent. Nous avons déjà obtenu une première retombée car cette bibliothèque constitue un excellent outil pour le prototypage d'applications de production ou de transformation d'Information Scientifique et Technique. Mais on peut souhaiter aller plus loin, afin d'obtenir un véritable Atelier de Génie Logiciel dédié à l'Information Scientifique et Technique ou plus exactement à son ingénierie. Nous démarrons nos réflexions sur ce sujet, et plusieurs pistes se dégagent.

Nous travaillons actuellement au niveau des outils de cet Atelier (dont la ILIB n'est qu'un produit cible). Pour les aspects "prototypage", la structure arborescente de SGML la rend particulièrement apte à l'utilisation d'outils d'intelligence artificielle et des travaux sur les "bonnes façons de manipuler des objets SGML en Lisp, Prolog ou C++" par exemple, sont les bienvenus ! Au niveau des spécifications, la définition formelle de SGML est un atout qu'il faut pouvoir utiliser. Pour la phase de génération, à coté de parsers SGML sophistiqués mais qui demandent de traiter un document de façon globale, il faudrait disposer d'outils permettant de travailler simplement sur un sous-ensemble, mais de façon plus conviviale que Lex ! Nous avons lancé la réalisation d'une maquette sur ce dernier thème.

Malgré la jeunesse du projet et le petit nombre d'intervenants, nous sommes déjà confrontés à des problèmes de maintenance. Les outils tels que make ne prennent pas toujours bien en compte la maintenance de bibliothèques de composants généraux ; ils privilégient l'assemblage de composants pour fabriquer des programmes et non les ensembles de composants, sans programme cible déclaré, mais avec des contraintes de cohérence. Les programmes réalisés sont souvent liés à une DTD, mais ce lien n'est jamais explicite. Pourtant, il faut maintenir la cohérence entre les programmes et les DTD. Cela veut dire que nous serons amenés à nous confronter à la gestion des objets, partie essentielle des AGL. Nous sommes encore trop peu avancés sur ce point pour "émettre des opinions définitives", mais il semble que notre problème se réduira probablement au choix d'un AGL existant et disposant de facilités de paramétrisation.

BIBLIOGRAPHIE

[BOR90] BORSTEIN J., RILEY V. "Hypertext Interchange Format", in Proceedings of the Hypertext Standardization Workshop, National Institute of Standards and Technology, 1990, pp 39-48

[CAL83] CALLON M., COURTIAL J-P., TURNER W.A., BAUIN S. "From Translation to Problematic Networks : An Introduction to Co-Word Analysis" in Social Science Information, vol. 22, 1983, pp 191-235

[COR91] CORET A. DUCLOY J. MENILLET D. "Les stations de travail des ingénieurs documentalistes à l'INIST" 9éme congrès IDT, Bordeaux, 1991, pp 189-195

[DAN90] DANIEL-VATONNE M.C. "Hypertextes : des principes communs et des variations" Technique et Science informatiques, 1990, Vol 9, No spécial : les hypertextes, pp 475-492

[DUC89] DUCLOY J. "L'INIST et ses choix technologiques pour l'informatisation" in Actes congrès INFORSID 89, Nancy 5/89, pp 139-145

[DUC91-1] DUCLOY J., GRIVEL L., LAMIREL J.C., POLANCO X., SCHMITT L. “INIST's Experience in Hyper-Document building from bibliographic Databases” Proceedings of RIAO’91 - Barcelone, April 91

[DUC91-2] DUCLOY J., LELU A. "Construction d'hyperdocuments à l'aide de procédés neuronaux" Génie Linguistique 91 - Versailles (FR) 16-17/01/1991

[DUS91] DUSOULIER N., DUCLOY J. "Processing of data and exchange of records in a scientific and technical information center. Formats : what for ?" UNIMARC/CCF Workshop, Florence (IT) (IFLA/UNESCO), 05-07 Juin 1991

[EC85] EC - FORMEX - "Formalized Exchange of Electronic Publications", Office for Official Publication in the European Communities, Luxembourg, 1985

[EWG90] European Workgroup on SGML : "MAJOUR (Modular Application for Journal)", STM : Scientific Technical and Medical Publishers, 1990.

[FRA90] FRANCOIS C. ”Analyse de références bibliographiques conformes à la norme ISO 2709 et conversion vers la norme SGML” Rapport de stage DESS Informatique, INIST - CNRS/ISIAL Université de Nancy 1, Nancy, 1990

[GRI91] GRIVEL L., LAMIREL J.C. "SDOC, a generation of hypertext structures" Proceedings of Multimedia Information Conference, Cambridge (UK), 15-18 juillet 1991

[GUI90] GUITTET J. “Combining CCF and SGML to exchange scientific and technical information” Proceedings of the first CCF Users Meeting - Unesco/IBE, Geneva, April 1989 (PGI-90/WS/4)

[HER88] HERWIJNEN E. "Practical SGML", Kluwer Academic Publishers, 1990

[IFL80] IFLA - “UNIMARC : Universal MARC Format” 2nd rev. ed. London : IFLA International Office for UBC, 1980

ISO 2709 - 1981. Format for Bibliographic Information Interchange on Magnetic Tape. In "Recueil de normes ISO 1, Documentation et information", 1988, ISO Organisation internationale de normalisation, Genève, pp 519-523

ISO 8879 - 1986. Information processing - Text and office systems - Standard Generalised Markup Language (SGML), 155 pages

[LEL90] LELU A. "Modèles neuronaux pour données textuelles" Journées ASU de statistiques - Tours (FR), 28 mai-1er juin 1990

[MEY90] MEYER B. "Conception et programmation par objets" Interedition - Paris, 1990

[MIC88] MICHELET B. “L’analyse des associations” Thèse de doctorat, Université de Paris VII, 1988

[NEW90] NEWCOMB S. ”X3V1.8MSD7, Journal of Development Standard Music Description Language” in Proceedings of the Hypertext Standardization Workshop, National Institute of Standards and Technology, 1990, pp179-188

[PGI88] UNESCO - PGI & UNISIST “CCF : The Common Communication Format - Second Edition” Paris, 1988 (PGI-88/WS/2)

[POL91] POLANCO X., SCHMITT L., BESAGNI D., GRIVEL L. "A la recherche de la diversité perdue : est-il possible de mettre en évidence les éléments hétérogènes d'un front de recherche?" Journées d'étude de la SFBA : les systèmes d'information élaborée - Ile Rousse (FR), 5-7 juin 1991

Résumé français : L'analyse de l'Information Scientifique et Technique (IST) stockée dans les bases de données bibliographiques requiert l'exploitation coordonnée de différentes techniques. Deux méthodes permettant de classer et représenter sur une carte thématique un ensemble de documents en se basant sur les mots-clés qui les indexent sont étudiées en profondeur. Ces études montrent que l'analyse et l'interprétation des résultats obtenus par de tels outils supposent un mélange d'exploration informelle intuitive et d'exploitation méthodique de l'information élaborée par ces outils d'analyse. En partant d'une métaphore, la navigation dans un océan d'informations, il est établi la nécessité de construire automatiquement des hypertextes à partir des données à analyser, en leur incorporant une carte de navigation et des indicateurs de positionnement thématique. L'exploration de cette voie débouche sur la conception et le développement d'un système informatique basé sur SGML (Standard Generalized MarkUp Language), HENOCH, qui permet de rassembler et d'organiser dans un SGBD (Système de Gestion de Bases de Données) des données bibliographiques normalisées et traitées par diverses techniques (linguistiques, classificatoires, cartographiques), puis de distribuer ces informations sur INTERNET via une interface de navigation générée automatiquement et adaptée à l'analyse de l'information. Il est montré expérimentalement que le couplage d'un hypertexte et d'un SGBD permet de modéliser et de mettre en place concrètement des mécanismes d'exploration de différentes représentations de l'information qui assistent l'utilisateur dans son interprétation des résultats des méthodes d'analyse. Les hypertextes générés par ce système sont évalué positivement par les utilisateurs de l'INIST-CNRS, où s'est effectué cette recherche. Ils en ont apprécié notamment l'ergonomie de navigation. Ses points faibles se situent au niveau du suivi des évolutions thématiques d'un corpus dans le temps**.** En guise de conclusion, quelques pistes d'améliorations sont ébauchées.

Titre anglais : Constructing hypertexts for the interpretation of scientific and technical information analysis methods

Résumé anglais : Analysis of Scientific and Technical Information (STI) from bibliographical databases requires the co-ordinated exploitation of various techniques. Two methods making it possible to classify and represent on a topic map a set of documents are studied in-depth. They are based on keywords indexing the documents. These studies show that the analysis and the interpretation of the results obtained by such tools require a mixture of intuitive browsing and of methodical exploration of the information worked out by these analysis tools. A metaphor, browsing in an ocean of information, highlights the necessity to generate automatically hypertexts based on the very data to be analysed and having their topic navigation map and some indicators of thematic position. This point leads to the design and the development of an information processing system (HENOCH), based on SGML (Standard Generalized MarkUp Language), to gather and organise in a DBMS (Data Base Management System) some bibliographical data which are standardised and treated by different techniques (computational linguistic, data analysis, clustering and mapping methods). Then this information is distributed on INTERNET via an interface of navigation generated automatically and adapted to the analysis of information. It is shown in experiments that the coupling of hypertext and database techniques is an appropriate way of organising such information when it is question of interpreting the results of some analysis methods. It makes it possible to model and to implement concretely the proper mechanisms of exploration of different representations. The hypertexts generated by this system are assessed positively by the users of the INIST-CNRS, where was carried out this search. They especially enjoy its ergonomics for navigating, while they feel some lacks for managing the comparison of different representations over time. As a conclusion, some tracks for improvements are outlined.

Dicipline : sciences de l'information et de la communication

Mots-clés : Veille scientifique, bibliométrie, infométrie, analyse de l'information, analyse de données, méthode des mots associés, classification, cartographie, hypertexte, internet, système de gestion de base de données.

Unité Recherche et Innovation, INIST-CNRS, 2 allée du Parc de Brabois, 54 514 Vandoeuvre-lès-Nancy Cedex, et
Centre de Recherche Rétrospective de Marseille (CRRM), Université Aix Marseille III
13 397 Marseille Cedex 20