

# Indexation « sémantique » de documents archéologiques

Aurélien BENEL. Sylvie CALABRETTO. Jean-Marie PINON \*  
INSA de Lyon / LISI / DAD

## Résumé

Notre étude traite de l'utilisation de « réseaux sémantiques de composition » comme langage pour l'indexation et la recherche de documents. Nous nous intéressons en effet au cas particulier d'une communauté de chercheurs en archéologie étudiant de brefs comptes-rendus de fouilles. Dans sa majeure partie, la problématique du projet s'apparente à celle de la recherche de documents. En effet, il s'agit de trouver un langage pivot entre l'utilisateur et l'ordinateur qui soit suffisamment puissant pour décrire le contenu sémantique des documents et des requêtes, mais aussi suffisamment formel pour que la machine puisse les mettre en correspondance. A cette problématique commune à un grand nombre de systèmes documentaires s'ajoutent certaines contraintes nouvelles. Tout d'abord, le problème n'est pas tant de trouver les documents que de les « retrouver », en effet les chercheurs lisent au moment de leur publication tous les articles susceptibles de les intéresser un jour, mais doivent être capables d'y accéder plus tard par rapport à un thème spécifique. Dans le même esprit, chaque chercheur ayant ses propres thèmes de recherche et même sa propre définition des termes du domaine, il serait illusoire de penser à une indexation effectuée par un tiers. L'idée que nous proposons est d'offrir aux chercheurs un outil de prise de notes de lecture et de recherche de ces notes, plutôt qu'un système documentaire soumis à une autorité extérieure qui les priverait de l'autonomie et de la liberté nécessaires à leur fonction. Une autre spécificité du problème réside dans la position centrale qu'occupent en archéologie l'espace et le temps, ces deux modalités si difficiles à modéliser et si souvent absentes des moteurs de recherche. Nous montrerons comment un langage documentaire basé sur la composition peut répondre aux attentes d'un tel projet et exprimer entre autres les relations spatio-temporelles ainsi que la place du sujet pensant.

## Mots-clefs

Bibliothèques virtuelles, Recherche de documents, Réseaux sémantiques, Taxinomies, Ordres partiels, Interactions Homme - Machine.

## Keywords

Virtual Libraries, Document Retrieval, Semantic Networks, Taxonomies, Partial Ordering, Computer-Human Interactions.

---

\* Aurelien.Benel@insa-lyon.fr, {Sylvie.Calabretto, Jean-Marie.Pinon}@lisi.insa-lyon.fr  
Bâtiment 502 – 20, avenue Albert Einstein – 69621 Villeurbanne Cedex  
Tél. : (33) 04 72 43 84 81– Fax : (33) 04 72 43 85 18

## Introduction

Ce travail s'inscrit dans une collaboration avec l'École Française d'Archéologie d'Athènes autour du projet de mise « en ligne » de la *Chroniques des fouilles* du *Bulletin de Correspondance Hellénique*, soit 80 ans d'archéologie en Grèce et à Chypre en de courts articles dont le nombre est de l'ordre de la dizaine de milliers. Dès la première phase du projet (cf. [BENE 98]), la position centrale que devrait occuper le « moteur de recherche » fut mise en évidence. Le fait qu'il s'agisse d'une part de documents destinés à des chercheurs et d'autre part de documents archéologiques rendait en grande partie inadéquats les systèmes documentaires actuels.

L'une de leurs limites selon nous est l'application du modèle issu de l'informatique de gestion, dans lequel le cadre des informations est fixé par la direction de l'entreprise, les informations sont saisies par des employés et sont consultées par des clients. Si ceci reste acceptable au niveau d'une bibliothèque « grand public », ceci est tout à fait inadapté dans le cas qui nous intéresse des documents pour experts. Dans une profession où la consultation de documents occupe une place centrale, imposer à l'expert une description des documents, c'est nier son expertise. En effet, la problématique change quelque peu, il ne s'agit pas tant de trouver des documents que de les « retrouver ». L'expert lit, lors de leur parution, la plupart des documents susceptibles de l'intéresser un jour, et, face à une question ultérieure, devra s'orienter dans son raisonnement par rapport à ses raisonnements précédents et à ceux de ses collègues. L'autre principale limite des systèmes actuels est leur difficulté à exprimer l'espace et le temps, ces deux modalités occupant une position centrale en archéologie.

Pour résoudre ce problème, nous proposons d'adopter une indexation « sémantique ». Dans une première partie nous définirons ce que nous entendons par « sémantique » et les recommandations qui en découlent pour notre système documentaire. Dans une seconde partie, nous proposerons un modèle basé sur la taxinomie et adoptant la forme d'un ordre partiel. Enfin, dans la dernière partie nous donnerons un aperçu d'un système documentaire basé sur un tel modèle.

## 1. Sémantique et activité documentaire

« Sémantique » : un terme à la mode ? Un but illusoire ? Non, à condition d'en choisir une définition adéquate. G. Mounin distingue trois théories de la sémantique en linguistique (cf. [MOUN 97]) : une théorie « *situationnelle* » dans laquelle la signification est donnée par la situation dans laquelle le locuteur s'exprime et la réponse qu'elle provoque chez l'auditeur, une théorie *contextuelle* dans laquelle le sens d'un mot s'obtient par la moyenne de ses emplois linguistiques, et enfin une théorie *structurelle* selon laquelle il existe une structuration hors contexte des termes soit sur des critères morphologiques soit par rapport aux concepts qu'ils représentent.

### a. Théorie structurelle

En ce qui concerne les applications de la théorie *structurelle* à la recherche d'information, nous renverrons le lecteur à [GENE 99]. Pour notre part, nous nous abstenons d'approfondir le sujet tant il nous paraît complexe. En effet, dans le cas où l'on choisit de décomposer les concepts en des concepts plus simple, la difficulté réside dans le choix d'un noyau de concepts indépendants les uns des autres et suffisants pour exprimer tous les autres. Depuis les *Catégories* d'Aristote, chaque nouvelle « ontologie » destinée à surpasser toutes les précédentes est venue les rejoindre dans leur incomplétude. Dans le cas où l'on renonce à une définition des concepts et l'on préfère un simple réseau

de relations (thesaurus...), la difficulté réside dans le choix de stratégies porteuses de sens pour réduire la complexité algorithmique qui découlerait de la transformation par transitivité de n'importe quel concept en n'importe quel autre.

### **b. Théorie contextuelle**

Pour ce qui est de la théorie *contextuelle*, on pense à tous les outils statistiques utilisés sur les textes intégraux. La tâche nous paraît des plus difficiles, en raison de l'absence de correspondance exacte entre les termes et les concepts qu'ils représentent. Si des raffinements toujours plus ingénieux permettent de contourner le polymorphisme des termes et la synonymie, l'homonymie quant à elle reste un obstacle de taille : qu'entend-on par « moyenne » sémantique des emplois d'un terme lorsque les concepts représentés n'ont rien de commun ? N'est-ce pas aussi malaisé que de résumer un dictionnaire ?

### **c. Théorie « situationnelle »**

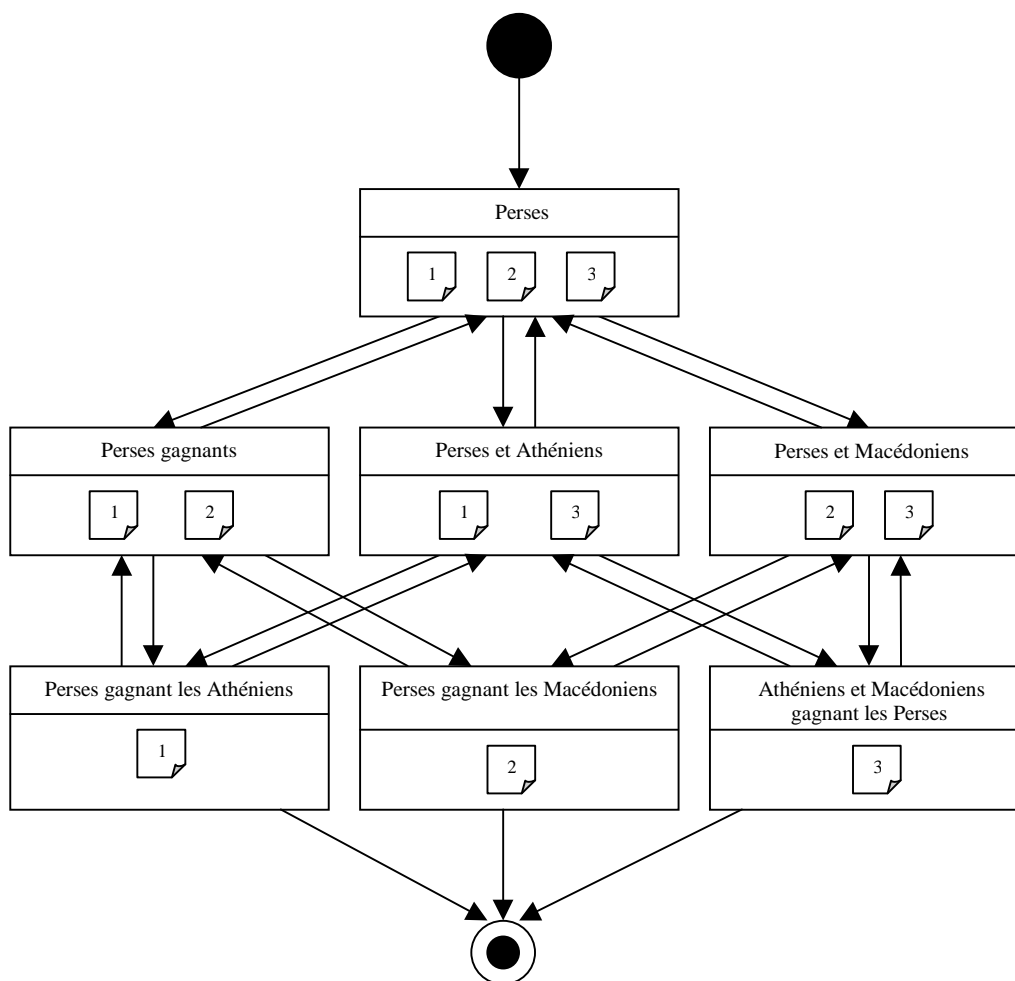
Concentrons nous maintenant sur la théorie « *situationnelle* ». Elle nous permet de déduire que l'indexation d'un document est dépendante entre autres de son indexeur (cf. [DAVI 95]), des personnes auxquelles il est destiné et des indexations déjà réalisées sur les autres documents du corpus. Cette dépendance ne doit non pas être considérée comme un défaut d'objectivité mais comme le réceptacle même du sens. Un certain nombre de chercheurs ont pris position pour que les interactions sociales ne soient pas oubliées dans les bibliothèques virtuelles (cf. [TOCH 94]). Pourquoi ? Il ne s'agit pas uniquement d'assurer l'équilibre psychologique des chercheurs, mais de garantir la qualité du travail en bibliothèque. En effet, la méthode optimale pour chercher des documents reste d'interroger (directement ou par l'intermédiaire de bibliographies, dossiers, index...) des spécialistes (bibliothécaires ou autres chercheurs). Eux seuls sont capables de comprendre le contenu des ouvrages, de porter dessus un jugement critique, d'en faire une synthèse, de faire des recoupements avec d'autres ouvrages, de les compléter par des informations non – officielles ou subjectives, de comprendre la question du chercheur d'information et de l'aider si besoin à raffiner ou à étendre sa requête. Un autre aspect important réside dans le travail de « capitalisation » et d'organisation de la connaissance que réalise le chercheur (sous forme de bibliographies, de fiches ou notes de lecture...). Nous proposons de placer ces deux composantes strictement humaines au centre de notre système documentaire. Reste maintenant à voir quelle pourrait être *l'aide*, limitée à ce qui peut être formalisé, d'un ordinateur dans un tel système.

N'est-ce pas ce que [BUSH 45], souvent considéré comme l'ouvrage fondateur de l'informatique documentaire et de l'hypertexte, recommandait pour son système de consultation des publications savantes ? Le lecteur devait pouvoir retrouver ses « chemins » de pensée empruntés dans le passé et suivre les chemins empruntés par d'autres (collègues, tuteurs, bibliothécaires). Et de plus, il s'agissait d'*assister* le lecteur de publications savantes en le soulageant des aspects répétitifs de son activité et en le laissant se concentrer sur les aspects créatifs, intuitifs et à haut niveau d'abstraction.

## **2. Proposition de modèle**

### **a. Taxinomie de documents**

Les processus d'indexation et de recherche d'un document dans un corpus peuvent être identifiés comme étant des processus de *définition* d'un document, *définition* au sens de *distinction*. Il s'agit en effet de distinguer un unique document parmi les documents du corpus. Ce processus de distinction d'un corpus élémentaire (ne contenant que le document)

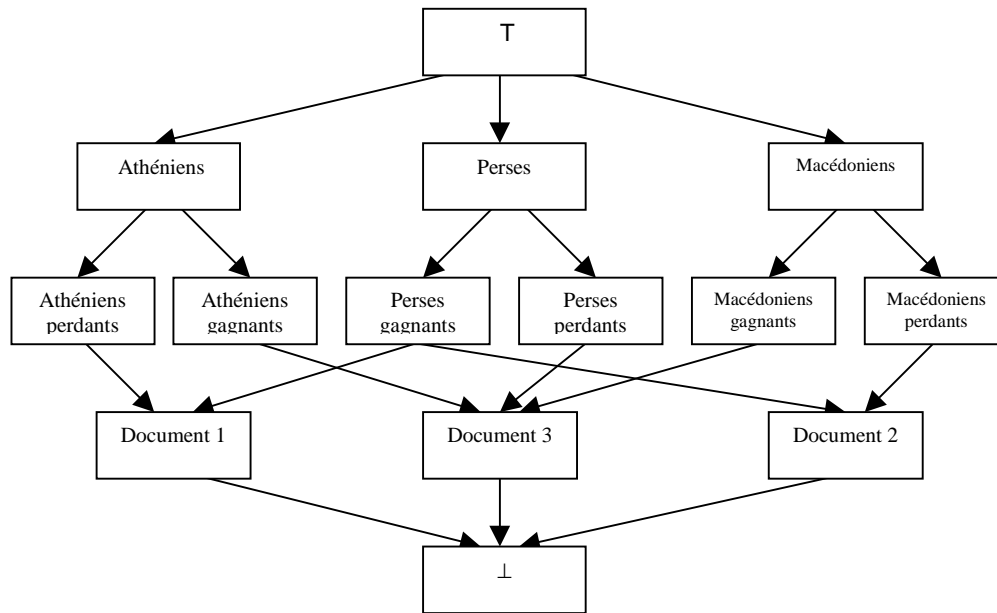


**Figure 1 – Exemple de recherche dans un corpus.**

par rapport au corpus de départ peut être décomposé en un ensemble de distinctions qui appliquées successivement forment des chemins à travers des corpus de plus en plus petits. La figure 1 illustre, pour un corpus d'exemple, les transitions (sous forme de flèches) et les sous-corpus possibles (sous forme de boîtes) de l'état initial (disque simple) à l'état final (disques concentriques). Le modèle des tâches « idéal » voudrait que chaque distinction se traduise par un saut vers un corpus contenant un document de moins et que tous les sous-corpus possibles soient présents. La complexité algorithmique qui en découlerait n'a en fait que peu de justifications, et nous adopterons plutôt, comme nous allons l'exposer plus loin un modèle mettant à profit l'intelligence humaine pour réaliser des classifications porteuses de sens et de complexité moindre.

### ***b. Taxinomie de descripteurs***

Nous proposons d'indexer les documents en construisant, « au-dessus » de leur identifiant, un ordre partiel de descripteurs (à rapprocher des hiérarchies de type dans [SOWA 92]). Cette entorse à l'habituelle structure arborescente permet de représenter un descripteur comme la composition d'autres descripteurs. Ce mécanisme s'il est intéressant pour des concepts ou des relations entre concepts est primordial pour des régions spatiales ou des intervalles temporels (cf. [PRED 99]).



**Figure 2 – Exemple d’index correspondant au corpus de la figure 1.**

### ***c. D’une taxinomie à l’autre***

Notre propos est maintenant de définir le lien existant entre la taxinomie des descripteurs et celle des documents. En effet trop de systèmes, quand ils permettent le raffinement des requêtes, autorisent l’ajout de n’importe quel descripteur sans faire de différence entre un descripteur qui serait incompatible avec le précédent, un autre qui en serait une spécialisation, ou un troisième qui formerait avec le précédent un nouveau descripteur.

Dans notre modèle, il s’agit d’associer un corpus à chaque sélection (simple ou multiple) de descripteurs. Prenons l’exemple des figures 1 et 2 : dans l’état initial le descripteur universel  $T$  est sélectionné (cf. Figure 2), le corpus correspondant est  $\{1,2,3\}$  (cf. Figure 1). Dans cette configuration, *Athéniens* et *Macédoniens* sont « possibles » : ils décrivent respectivement  $\{1,3\}$  et  $\{2,3\}$  et il existe une transition directe du corpus actuel  $\{1,2,3\}$  vers ces corpus (cf. Figure 1). *Perses* est « connu » : il décrit le corpus actuel  $\{1,2,3\}$ . *Perses gagnants* et *Perses perdants* sont possibles. Supposons que l’on sélectionne alors *Perses gagnants*, le corpus correspondant est  $\{1,2\}$ , *Perses* devient implicite, *Perses perdants* ainsi que *Athéniens gagnants* et *Macédoniens gagnants* sont impossibles, *Document 1* et *Document 2* sont possibles, *Athéniens perdants* et *a fortiori Athéniens* sont implicites à *Document 1*, *Macédoniens perdants* et *a fortiori Macédoniens* sont implicites à *Document 2*. Supposons maintenant que l’on sélectionne *Macédoniens perdants*, le corpus correspondant est  $\{2\}$ , *Document 2* est connu, *Macédoniens* est implicite, *Athéniens* est impossible et *a fortiori Athéniens gagnants*, *Athéniens perdants*, *Document 1* et *Document 3*.

Si cet exemple avec trois documents vous a semblé laborieux, imaginez un cas réel avec des dizaines de milliers de documents. Or, le raisonnement fait ci-dessus est purement formel. L’acte « créatif » a été effectué par l’indexeur et par la personne qui a sélectionné les descripteurs et non pas dans ce « raisonnement ». C’est justement pour ce raisonnement formel que l’ordinateur peut nous apporter une aide.

Un souci légitime pourrait porter sur la complexité temporelle des calculs (ceux-ci sont en effet basés sur la fermeture transitive du graphe acyclique) et par là leur inadéquation à une interface homme-machine. Cependant, une solution inspirée de [AITK 89] revient à

calculer à l'avance et à stocker la fermeture transitive du graphe pour ensuite, au moment de l'interrogation, réduire les calculs à un petit nombre d'opérations évoluant peu ou pas avec le nombre de documents.

### 3. Préfiguration du système

Une première classe d'informations documentaires est de nature objective, explicite, et incontestable. Il s'agit entre autres de la référence logique et de la référence physique des documents. Dans le cas où notre document est un article de périodique, la référence physique est constituée du nom de la série, de son numéro et des pages qui contiennent l'article (et qui peuvent en contenir d'autres). La référence logique quant à elle est composée de la hiérarchie des titres. On peut noter que la nature de ces données ne nécessite qu'une structure arborescente.

Une deuxième classe d'informations documentaires concerne ce qui est subjectif et sujet à discussion. Nous la nommerons référence sémantique. Celle-ci sera composée des indexations personnelles, complémentaires voire contradictoires. A chacune de ces indexations sera attaché son auteur. Dans ce contexte, l'indexation « officielle » (de la bibliothèque ou du périodique) est une indexation personnelle réalisée par un individu dans le cadre d'une mission pour une organisation. En plus de cette dimension du sujet (organisée selon les écoles de pensée, les organisations, etc.), d'autres dimensions sont à prendre en compte suivant le domaine : dans le cas de l'archéologie, par exemple, ce seront les artefacts, l'espace et le temps.

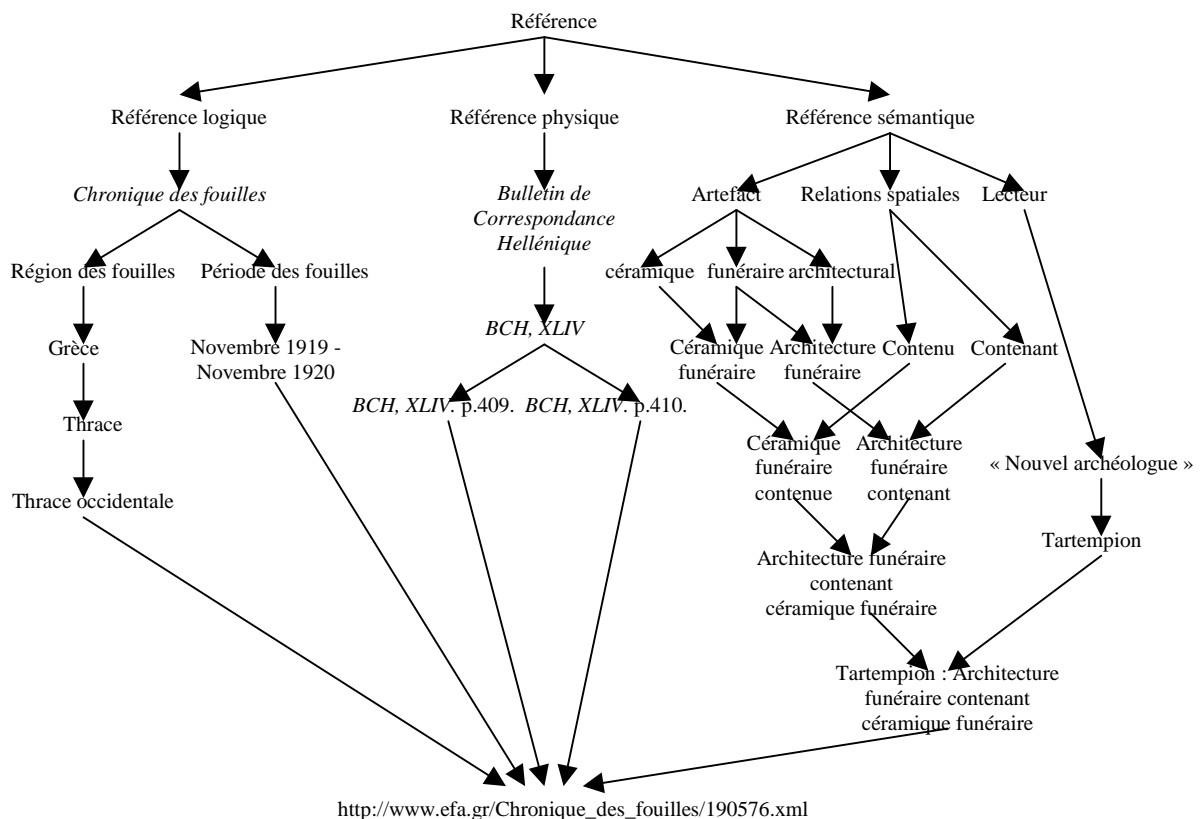


Figure 3 – Exemple d'index "sémantique" d'un document archéologique.

## Conclusion

Nous avons proposé un système documentaire basé sur des descripteurs informels insérés dans une structure formelle de composition. Ce compromis entre formel et informel laisse un maximum de latitude au chercheur tout en l'aidant à s'orienter dans le corpus en mettant en évidence les dépendances entre descripteurs. La généralité du modèle et des traitements associés permet de les appliquer à des descripteurs de type très variés : concepts, relations conceptuelles, intervalles de temps, régions spatiales, personnes...

Nous affirmons que le système ainsi conçu est « sémantique » (au sens de la théorie « situationnelle » de la sémantique selon G. Mounin) en affectant l'ordinateur aux tâches répétitives et en laissant l'être humain se concentrer sur les aspects créatifs, intuitifs et à haut niveau d'abstraction de son activité.

## Bibliographie

- [AITK 89] **AÏT-KACI H., et al.** Efficient Implementation of Lattice Operations. In: *ACM Transactions on Programming Languages and Systems, Vol. 11, No 1 (Jan. 1989)*. p.115-146.
- [BENE 98] **BENEL A.** *La Chronique des fouilles : de la bibliothèque à l'Internet. Etude globale du projet.* Rapport interne. Ecole Française d'Athènes, 1998.
- [BUSH 45] **BUSH V.** As We May Think. In: *The Atlantic Monthly*. July 1945.
- [DAVI 95] **DAVID C., et al.** Indexing as Problem Solving: a Cognitive Approach to Consistency. In: *ACSI'95: Annual Conference of the Canadian Association for Information Science, Edmonton (Alberta), June 7-10, 1995*.
- [GENE 99] **GENEST D.** Vers un système de recherche documentaire basé sur les graphes conceptuels. In: *Actes du XVII<sup>e</sup> congrès INFORSID. La Garde, France, 2-4 juin 1999*. p.115-131.
- [MOUN 97] **MOUNIN G.** *La sémantique.* Réédition revue et corrigée de l'édition de 1972. Editions Payot, 1997.
- [PRED 99] **PREDIGER S., WILLE R.** The Lattice of Concept Graphs of a Relationally Scaled Context. To be published in: *Knowledge Science and Engineering with Conceptual Structures, Lecture Notes in Artificial Intelligence*, Springer, Berlin, 1999.
- [SOWA 92] **SOWA J. F.** *Semantic Networks.* IBM Systems Research, 1992.
- [TOCH 94] **TOCHTERMANN K.** A First Step Toward Communication in Virtual Libraries. *Digital Libraries '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station (Texas), June 19-21 1994*.