

Aurélien Bénel
DEA Informatique de Lyon
Année 1998-1999

Mémoire de DEA

Indexation « sémantique » de documents pour experts

Le cas de la *Chronique des fouilles* de l'Ecole Française d'Athènes

Stage effectué au sein
de l'équipe Documentation et Aide à la Décision
du Laboratoire d'Ingénierie des Systèmes d'Information
de l'INSA de Lyon
sous la responsabilité de Sylvie Calabretto
en collaboration avec l'Ecole Française d'Archéologie d'Athènes

Remerciements

Je tiens à remercier tous ceux qui ont rendu possible ce travail et plus particulièrement : Jacqueline Martinez du Département Informatique de l'INSA de Lyon, Sylvie Calabretto et Jean-Marie Pinon du Laboratoire d'Ingénierie des Systèmes d'Information, Andréa Iacovella et Roland Etienne de l'Ecole Française d'Athènes.

Résumé

Notre étude traite de l'utilisation de taxinomies 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.

Table des matières

1	PROBLÉMATIQUE	1
2	SÉMANTIQUE ET ACTIVITÉ DOCUMENTAIRE : ETAT DE L'ART	3
2.1	THÉORIE STRUCTURELLE.....	3
2.2	THÉORIE CONTEXTUELLE	4
2.3	THÉORIE « SITUATIONNELLE »	4
3	PROPOSITION DE MODÈLE	6
3.1	PRÉSENTATION INFORMELLE	6
3.1.1	<i>Taxinomie de documents</i>	6
3.1.2	<i>Taxinomie de descripteurs</i>	7
3.1.3	<i>D'une taxinomie à l'autre</i>	8
3.2	SPÉCIFICATION FORMELLE DU TYPE TAXINOMIE	11
3.2.1	<i>Motivations</i>	11
3.2.2	<i>Pré-requis</i>	11
3.2.3	<i>Domaines de définition et d'application</i>	11
3.2.4	<i>Axiomatique</i>	13
3.3	CONSIDÉRATIONS PRAGMATIQUES	16
4	VALIDATION ET ÉTUDE CRITIQUE	17
4.1	VALIDITÉ PAR RAPPORT AUX BESOINS.....	17
4.1.1	<i>Exemple d'indexation de la Chronique des fouilles</i>	17
4.1.2	<i>Irréductibilité à la formalisation ?</i>	19
4.1.3	<i>Objectivité et Intersubjectivité</i>	19
4.1.4	<i>Typologie, Espace, et Temps</i>	20
4.2	VALIDATION OPÉRATOIRE	21
5	BILAN ET PERSPECTIVES	22
6	BIBLIOGRAPHIE	23
6.1	INFORMATIQUE, SCIENCES DE L'INFORMATION.....	23
6.2	ARCHÉOLOGIE.....	25
7	ANNEXE : PROTOTYPE EN PROLOG	26

1 Problématique

Ce travail s'inscrit dans une collaboration avec l'Ecole 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 » des documents pour experts. Dans une première partie nous

définirons ce que nous entendons par « sémantique » et à la lumière de cette définition, nous établirons un état de l'art sur l'indexation « sémantique » de documents ainsi que des recommandations 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 et en ferons une étude critique en termes de validité.

2 Sémantique et activité documentaire : Etat de l'art

« 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.

2.1 Théorie structurelle

En ce qui concerne les applications de la théorie *structurelle* à la recherche d'information, nous renverrons le lecteur à [GENE 99], [FOUR 98], [MECH 95], [OUNI 98]. Pour notre part, nous nous abstenons d'approfondir le sujet tant il nous paraît périlleux. En effet, dans le cas où l'on choisit de décomposer les concepts en des concepts plus simples, 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 » (cf. [LEHM 94]) 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 (cf. thesaurus Rameau dans [GENE 99]), 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.

2.2 Théorie contextuelle

Pour ce qui est de la théorie *contextuelle*, on pense à tous les outils statistiques utilisés sur les textes intégraux (cf. [FUTR 94], [LESK 96]). La tâche nous paraît des plus difficiles, en raison de l'absence de correspondance exacte entre les termes et les concepts qu'ils représentent (cf. [CHIA 99]). 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 ?

2.3 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 V. Bush (cf. [BUSH 45]), souvent considéré comme le 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). 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.

3 Proposition de modèle

3.1 *Présentation informelle*

3.1.1 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* (à l'aide de descripteurs par exemple) 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) 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.

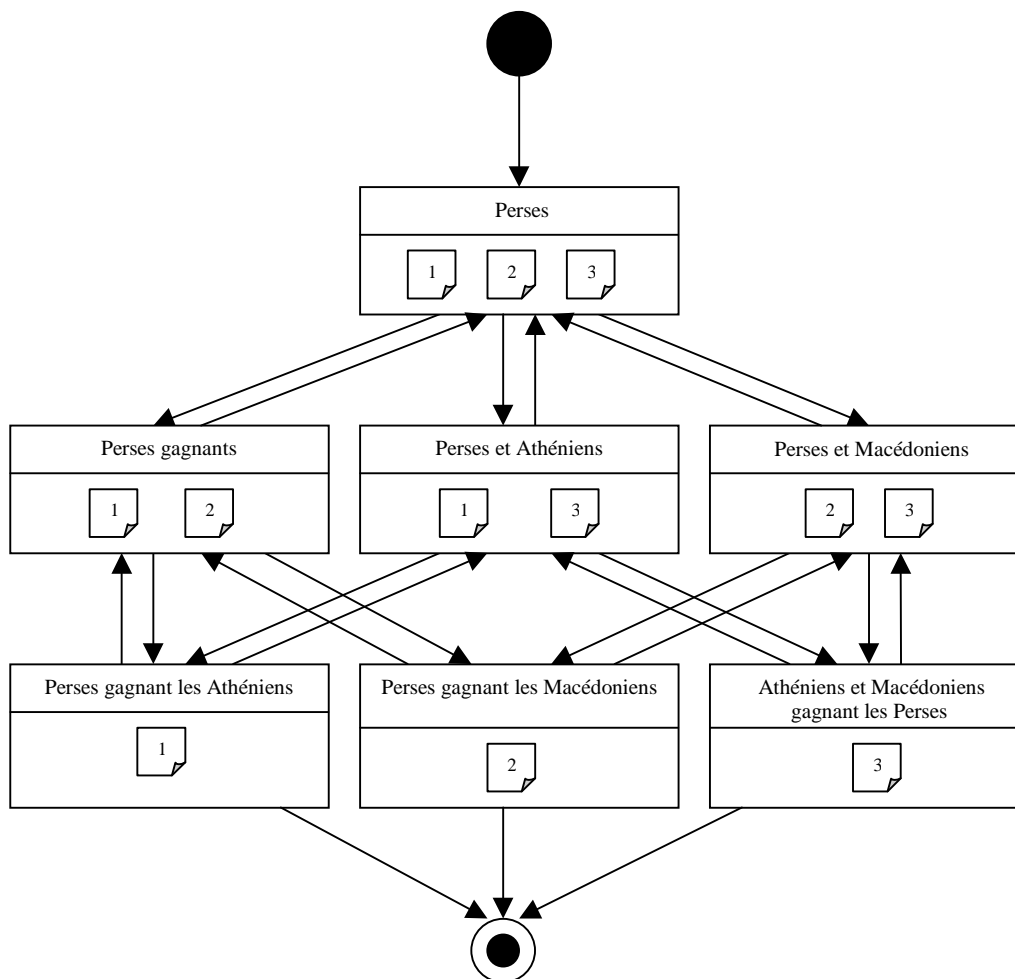


Figure 1 – Exemple de recherche dans une taxinomie de documents.

3.1.2 Taxinomie de descripteurs

Nous proposons d’indexer les documents en construisant, « au-dessus » de leur identifiant, un ordre partiel de descripteurs (à rapprocher dans [SOWA 92] de la relation « sorte de »). Cette entorse à l’habituelle structure arborescente permet de représenter un descripteur comme la composition d’autres descripteurs (cf. Figure 2). 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]).

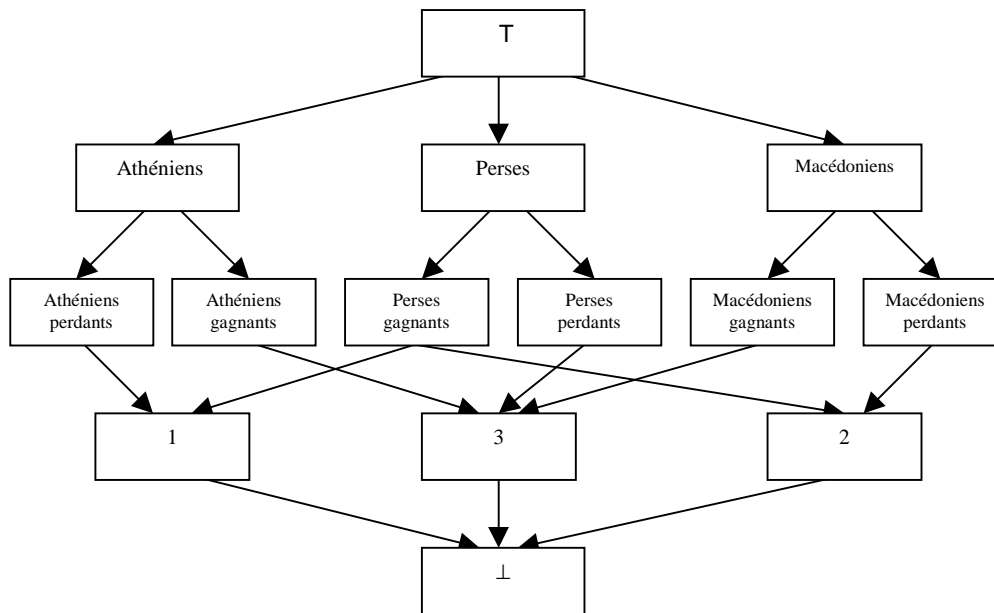


Figure 2 – Taxinomie de descripteurs pour l'exemple de la figure 1.

3.1.3 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.

La visualisation de ce « lien », utilisé comme support des interactions homme-machine, peut prendre plusieurs formes suivant que l'on choisit de se placer plutôt du côté des corpus ou des descripteurs. Un bon compromis pour éviter à la fois la surcharge cognitive et la désorientation consiste à se baser sur la taxinomie des descripteurs (le chercheur, indexeur potentiel,

doit se familiariser avec la structure de l'index) et à la « filtrer » à l'aide des relations entre corpus.

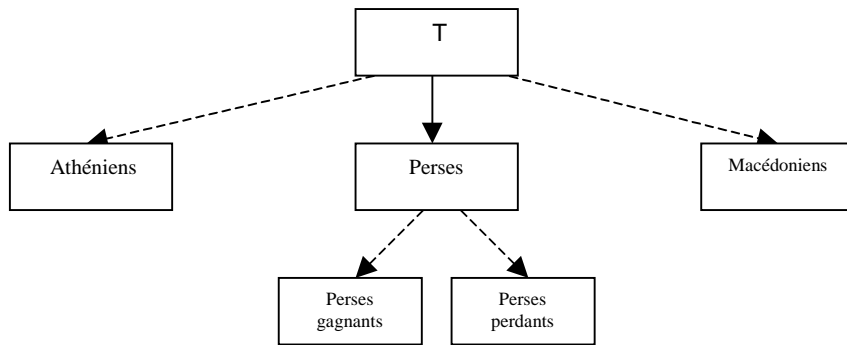


Figure 3 – Etat {1, 2, 3} de la consultation de l'index de la figure 2.

Prenons l'exemple des figures 1 et 2. L'état initial correspond à celui où le descripteur universel T est sélectionné (cf. Figure 3), c'est à dire pour lequel le corpus correspondant est $\{1,2,3\}$. Dans cette configuration, *Perses* est « connu » puisqu'il décrit le corpus $\{1,2,3\}$ qui généralise (au sens large) le corpus actuel. *Athéniens* est « possible » car il décrit le corpus $\{1,3\}$ pour lequel il existe une spécialisation (non vide) commune avec $\{1,2,3\}$ (cf. Figure 1). *Macédoniens*, *Perses gagnants* et *Perses perdants* sont également possibles.

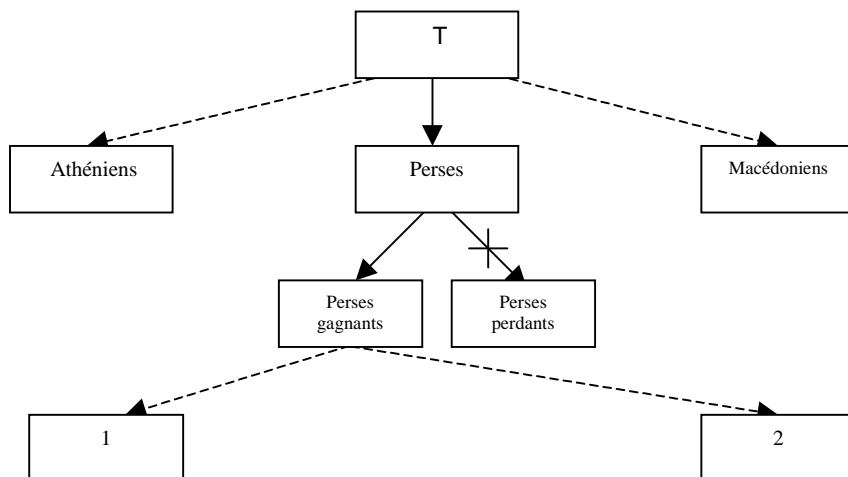


Figure 4 – Etat {1, 2} de la consultation de l'index de la figure 2.

Supposons que l'on sélectionne alors *Perses gagnants* (cf. Figure 4), le corpus correspondant est $\{1,2\}$. *Perses gagnants* est connu et *a fortiori* *Perses* et T . *Document 1* et *Document 2* sont possibles, ainsi que *Athéniens* et *Macédoniens*. *Perses perdants* est « impossible ». En effet il décrit le corpus $\{3\}$ or $\{3\}$ ne contient pas $\{1,2\}$ et il n'existe pas de spécialisation (non vide) commune à $\{3\}$ et à $\{1,2\}$ (cf. Figure 1).

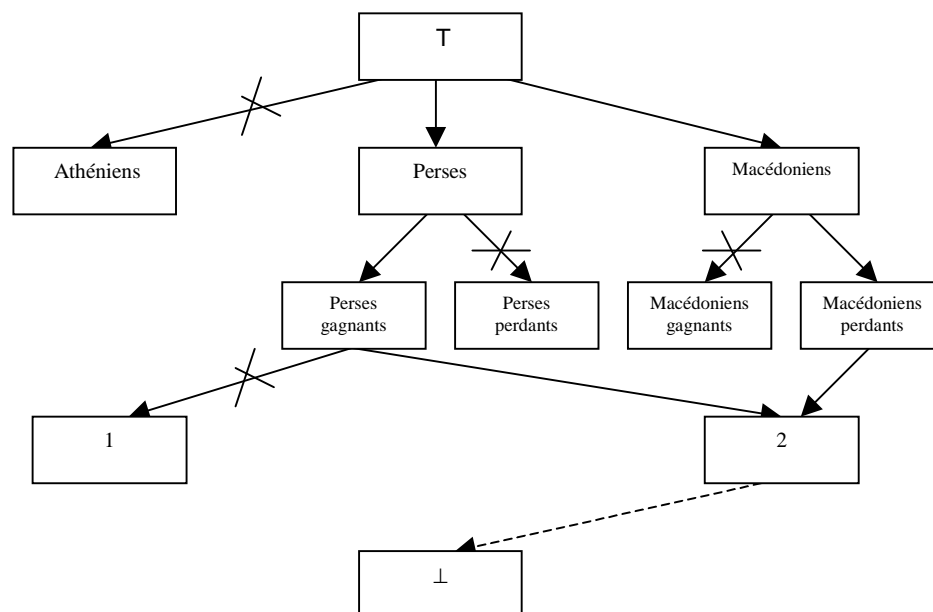


Figure 5 – Etat {2} de la consultation de l'index de la figure 2.

Supposons que l'on sélectionne ensuite *Macédoniens* (cf. Figure 5), le corpus correspondant est $\{2\}$, *Document 2* et *a fortiori* *Macédoniens perdants*, *Perses gagnants*, *Perses*, *Macédoniens* et T sont connus, *Athéniens* est impossible ainsi que *Macédoniens gagnants* et *Document 1*.

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.

3.2 Spécification formelle du type Taxinomie

3.2.1 Motivations

Dans la partie précédente, nous avons mis en évidence quelle part de l'activité documentaire pouvait être formalisée, à savoir la mise en évidence des dépendances entre descripteurs par rapport à un corpus. Dans cette partie, nous allons spécifier formellement ce que nous entendons par *arcs possibles*, *arcs impossibles*, *arcs connus* et *sélection d'un nouveau corpus*. Cette spécification prendra la forme de définitions de fonctions et de règles de réécriture.

3.2.2 Pré-requis

Pour éviter d'alourdir inutilement la spécification, on supposera prédéfinis les types *Booléen*, *Atome* et *Ensemble*. Dans la suite l'énonciation d'un booléen sera équivalente à son égalité à vrai. Les comportements des quantificateurs, des opérations booléennes, ensemblistes, ainsi que ceux des opérations sur les atomes sont les comportements usuels.

3.2.3 Domaines de définition et d'application

Les conventions utilisées (cf. [CALA 93]) veulent que la description de la « signature » d'une fonction se compose à gauche des « deux points » : du symbole de la fonction, de ses paramètres d'entrée représentés chacun par un « blanc souligné » (indiquant à la fois leur nombre et leur position), et à droite : du domaine d'entrée (représenté par un produit cartésien), d'une flèche et du domaine de sortie.

Fonction « constructeur de taxinomie »

$$\tau_0 : \emptyset \rightarrow \text{Taxinomie}$$

Fonction « constructeur de taxinomie par insertion d'un descripteur dans une taxinomie »

$$\tau (_ , _ , _ , _) : \text{Taxinomie} \times \text{Atome}^3 \rightarrow \text{Taxinomie}$$

Fonction « a pour fils dans une taxinomie »

$$_ \blacktriangleright _ : \text{Atome} \times \text{Taxinomie} \times \text{Atome} \rightarrow \text{Booléen}$$

Fonction « a pour descendant dans une taxinomie »

$$_ \blacktriangleright\blacktriangleright _ : \text{Atome} \times \text{Taxinomie} \times \text{Atome} \rightarrow \text{Booléen}$$

Fonction « a pour terminal dans une taxinomie »

$$_ \blacktriangleright\blacktriangleright\blacktriangleright _ : \text{Atome} \times \text{Taxinomie} \times \text{Atome} \rightarrow \text{Booléen}$$

Fonction « a pour corpus dans une taxinomie »

$$\bullet _ : \text{Taxinomie} \times \text{Atome} \rightarrow \text{Ensemble}$$

Fonction « descripteurs connus d'un corpus dans une taxinomie »

$$\checkmark _ : \text{Taxinomie} \times \text{Ensemble} \rightarrow \text{Ensemble}$$

Fonction « descripteurs possibles d'un corpus dans une taxinomie »

$$\square _ : \text{Taxinomie} \times \text{Ensemble} \rightarrow \text{Ensemble}$$

Fonction « descripteurs impossibles d'un corpus dans une taxinomie »

$$\boxtimes _ : \text{Taxinomie} \times \text{Ensemble} \rightarrow \text{Ensemble}$$

Fonction « arcs connus d'un corpus dans une taxinomie »

$$\checkmark _ : \text{Taxinomie} \times \text{Ensemble} \rightarrow \text{Ensemble}$$

Fonction « arcs possibles d'un corpus dans une taxinomie »

$$? _ : \text{Taxinomie} \times \text{Ensemble} \rightarrow \text{Ensemble}$$

Fonction « arcs impossibles d'un corpus dans une taxinomie »

$$\times _ : \text{Taxinomie} \times \text{Ensemble} \rightarrow \text{Ensemble}$$

Fonction « corpus demandé par sélection d'un descripteur à partir d'un corpus dans une taxinomie »

$\odot_{-} (_ , _) : \text{Taxinomie} \times \text{Ensemble} \times \text{Atome} \rightarrow \text{Ensemble}$

3.2.4 Axiomatique

La taxinomie élémentaire τ_0 est telle que pour cette taxinomie, le descripteur universel \top a pour fils le descripteur absurde \perp .

$$(\top \triangleright_{\tau_0} \perp) = \text{VRAI}$$

Pour toute taxinomie α , la relation *descendant* est obtenue par fermeture transitive de la relation *fils*.

$$(\forall \alpha \in \text{Taxinomie}) (\forall (x, y) \in \text{Atome}^2) \\ ((x \triangleright_{\alpha} y) \Rightarrow (x \blacktriangleright_{\alpha} y) = \text{VRAI})$$

$$(\forall \alpha \in \text{Taxinomie}) (\forall (x, y, z) \in \text{Atome}^3) \\ ((x \triangleright_{\alpha} y) \wedge (y \blacktriangleright_{\alpha} z)) \Rightarrow x \blacktriangleright_{\alpha} z = \text{VRAI}$$

On définit une taxinomie non élémentaire en « insérant » dans une autre taxinomie un descripteur entre un descripteur et un de ses descendants.

$$(\forall \alpha \in \text{Taxinomie}) (\forall (x, y, z) \in \text{Atome}^3) \\ ((x \blacktriangleright_{\alpha} z) \Rightarrow ((x \triangleright_{\tau(\alpha, x, y, z)} y) \wedge (y \triangleright_{\tau(\alpha, x, y, z)} z)) = \text{VRAI})$$

$$(\forall \alpha \in \text{Taxinomie}) (\forall (x, y, z, r, s) \in \text{Atome}^5) \\ (((r \triangleright_{\alpha} s) \wedge ((r \neq x) \vee (s \neq z))) \Rightarrow (r \triangleright_{\tau(\alpha, x, y, z)} s) = \text{VRAI})$$

Est appelé *terminal d'un descripteur* tout descripteur atteignable du premier par un chemin (éventuellement vide) de filiation et qui a pour fils le descripteur absurde.

$$(\forall \alpha \in \text{Taxinomie}) (\forall x \in \text{Atome}) \\ ((x \triangleright_{\alpha} \perp) \Rightarrow (x \blacktriangleright_{\alpha} x) = \text{VRAI})$$

$$(\forall \alpha \in \text{Taxinomie}) (\forall (x, y, z) \in \text{Atome}^3)$$

$$((x \triangleright_{\alpha} y) \wedge (y \triangleright_{\alpha} z)) \Rightarrow (x \triangleright_{\alpha} z = \text{VRAI})$$

On appelle *corpus d'un descripteur*, l'ensemble des terminaux de ce descripteur.

$$(\forall \alpha \in \text{Taxinomie}) (\forall x \in \text{Atome}) \\ (\bullet_{\alpha} x = \{ y \in \text{Atome} \mid (x \triangleright_{\alpha} y) \})$$

L'ensemble des descripteurs connus d'un corpus est l'ensemble des descripteurs dont le corpus contient (au sens large) le corpus en question.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) \\ (\boxtimes_{\alpha} L = \{ x \in \text{Atome} \mid (\bullet_{\alpha} x \supseteq L) \})$$

L'ensemble des descripteurs impossibles d'un corpus est l'ensemble des descripteurs dont le corpus forme une intersection non vide avec le corpus en question ou pour lesquels l'un des deux corpus est vide.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) \\ (L \neq \emptyset \Rightarrow \boxtimes_{\alpha} L = \{ x \in \text{Atome} \mid (\bullet_{\alpha} x \neq \emptyset) \wedge (\bullet_{\alpha} x \cap L \neq \emptyset) \})$$

L'ensemble des descripteurs possibles d'un corpus est l'ensemble des descripteurs fils qui ne sont ni connus ni impossibles pour le corpus en question.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) \\ (\boxdot_{\alpha} L = \{ x \in \text{Atome} \mid (y \triangleright_{\alpha} x) \wedge (x \notin \boxtimes_{\alpha} L) \wedge (x \notin \boxtimes_{\alpha} L) \})$$

L'ensemble des arcs connus d'un corpus est l'ensemble des couples de descripteurs père-fils dont le père et le fils sont connus pour le corpus en question.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) \\ (\checkmark_{\alpha} L = \{ (x, y) \in \text{Atome}^2 \mid (x \in \boxtimes_{\alpha} L) \wedge (x \triangleright_{\alpha} y) \wedge (y \in \boxtimes_{\alpha} L) \})$$

L'ensemble des arcs possibles d'un corpus est l'ensemble des couples de descripteurs père-fils dont, pour le corpus en question, le père est connu et le fils est possible.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) \\ (?_{\alpha} L = \{ (x, y) \in \text{Atome}^2 \mid (x \in \boxplus_{\alpha} L) \wedge (x \triangleright_{\alpha} y) \wedge (y \in \square_{\alpha} L) \})$$

L'ensemble des arcs impossibles d'un corpus est l'ensemble des couples de descripteurs père-fils dont, pour le corpus en question, le père est connu et le fils est impossible.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) \\ (\star_{\alpha} L = \{ (x, y) \in \text{Atome}^2 \mid (x \in \boxplus_{\alpha} L) \wedge (x \triangleright_{\alpha} y) \wedge (y \in \boxtimes_{\alpha} L) \})$$

En *sélectionnant* un descripteur connu pour un corpus, on étend le corpus de recherche à l'union de ce corpus et du corpus du descripteur.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) (\forall x \in \text{Atome}) \\ ((x \in \boxplus_{\alpha} L) \Rightarrow (\odot_{\alpha} (L, X) = (L \cup \bullet_{\alpha} x)))$$

En *sélectionnant* un descripteur possible pour un corpus, on restreint le corpus de recherche à l'intersection de ce corpus et du corpus du descripteur.

$$(\forall \alpha \in \text{Taxinomie}) (\forall L \in \text{Ensemble}) (\forall x \in \text{Atome}) \\ ((x \in \square_{\alpha} L) \Rightarrow (\odot_{\alpha} (L, X) = (L \cap \bullet_{\alpha} x)))$$

3.3 Considérations pragmatiques

Un souci légitime pourrait porter sur la complexité temporelle des calculs et par là leur inadéquation aux interactions homme-machine. En y regardant de plus près, les calculs ci-dessus consistent, à déduire récursivement une relation et à faire des opérations élémentaires sur des ensembles. On pourrait envisager de calculer à l'avance et de stocker tous ces calculs, cependant le comportement optimal entre coût en espace de stockage et coût en temps de calcul revient à calculer à l'avance la relation obtenue par récursivité et à faire « à la volée » les opérations ensemblistes (tâches dont les systèmes de gestion de bases de données s'acquittent particulièrement bien). Cette solution est inspirée de celle de H. Aït-Kaci (cf. [AITK 89]) qui, pour calculer efficacement le minimum local de deux nœuds dans un treillis, associe à chaque nœud l'ensemble (sous forme d'un vecteur de booléens) de ses minima non absurdes. Ainsi, au moment de l'interrogation, les calculs sont réduits à un petit nombre d'opérations évoluant peu ou pas avec le nombre de documents.

4 Validation et étude critique

4.1 Validité par rapport aux besoins

Si l'acceptation de notre article (cf. [BENE 99]) au colloque du chapitre français de l'ISKO, nous permet d'augurer de la validité de notre travail par rapport aux besoins de la consultation des documents pour experts, nous allons toutefois étudier, pour la cas spécifique de l'archéologie, l'adéquation de notre modèle aux besoins exprimés par les ouvrages méthodologiques.

Dans un premier temps, nous proposerons une application de notre modèle pour indexer un article de la *Chronique des fouilles*. Ensuite, nous ferons un tour d'horizon des quelques besoins fondamentaux de l'archéologie et nous évaluerons l'adéquation ou la non adéquation de notre modèle à ces besoins.

4.1.1 Exemple d'indexation de la *Chronique des fouilles*

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.

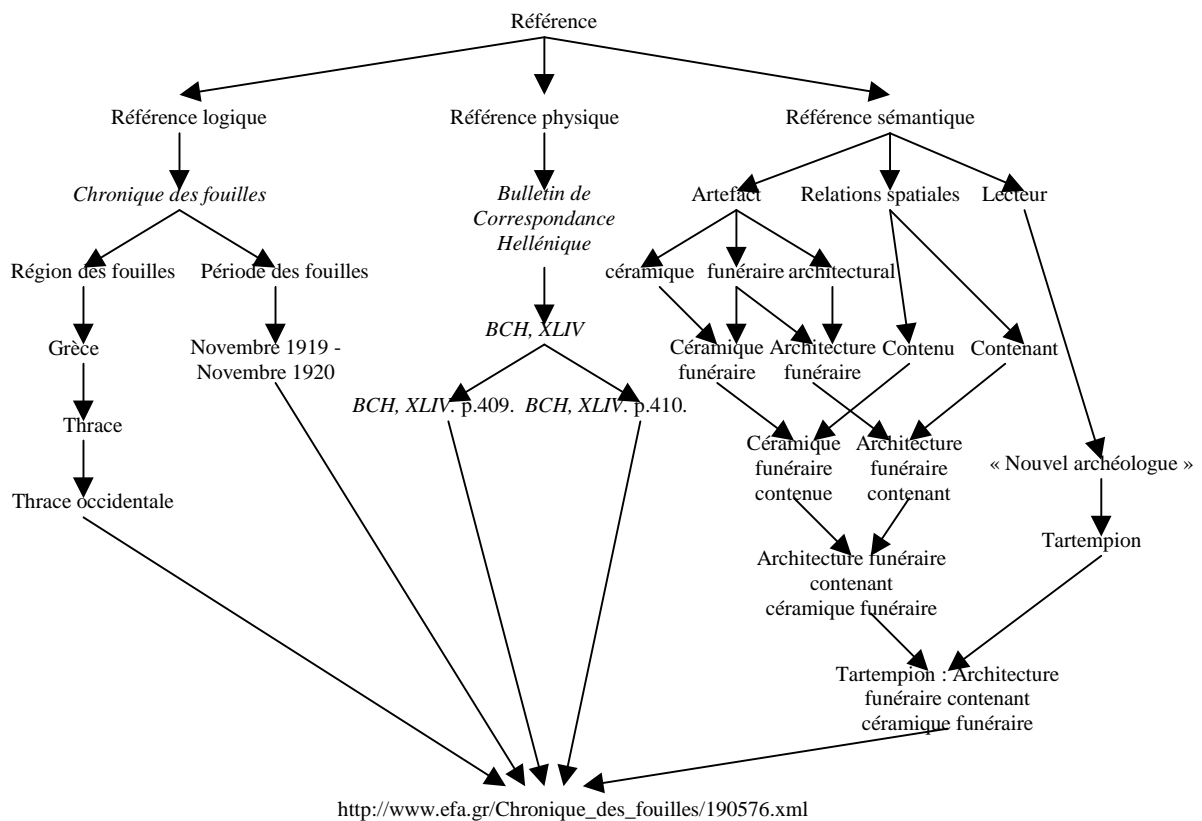


Figure 6 – Exemple d’index "sémantique" d’un article de la *Chronique des fouilles*.

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.

4.1.2 Irréductibilité à la formalisation ?

Le problème de la formalisation en archéologie est un sujet particulièrement polémique. A la « Nouvelle Archéologie » qui voulait soumettre l'archéologie à la validation, et à Gardin par exemple qui prônait le remplacement du contenu textuel des publications en archéologie par des propositions logiques (cf. [GARD 86]), de nombreux archéologues « traditionnels » se sont opposés (cf. [COUR 82]) pour défendre l'appartenance de l'archéologie aux Sciences Humaines. J.-C. Gardin lui-même reconnaît (cf. [GARD 86]) d'ailleurs le peu d'intérêt que rencontrèrent ses projets de « banques de données archéologiques » et interprète ces échecs comme étant dus à la difficulté de distinguer en archéologie les « faits », des conclusions ou des interprétations.

Sans vouloir rentrer dans le débat qui oppose les archéologues, nous pouvons toutefois apprécier *a posteriori* nos précautions épistémologiques de réduire la part formelle de notre système à de simples relations de composition et d'avoir refusé de définir un langage nomologique.

4.1.3 Objectivité et Intersubjectivité

Comme le souligne J.-C. Gardin (cf. [GARD 84]), on ne peut parler de bases de données archéologiques sans s'interroger sur la reconnaissance de ces données par l'ensemble de la communauté scientifique en archéologie : « Par quels mécanismes obtiendra-t-on que l'accord initial [...] engageant une population de chercheurs limitée dans l'espace et le temps, s'étende ensuite de façon quasi-statutaire [...] ? ». Plutôt que d'espérer en un hypothétique consensus assurant l'objectivité des données, R. Ginouves (cf. [GINO 78]) conseille de viser l'intersubjectivité. Ce point de vue est adopté dans P. Desfarges (cf. [DESF 91]) : « Les objets n'ont pas d'attributs par eux mêmes mais par leurs sources » et mis en pratique dans le système

FRANTIQ, dans lequel sont enregistrés des « discours » d’auteurs sur des artefacts, et non des données impersonnelles.

Sur ce point également, notre système répond aux attentes des archéologues en exprimant l’intersubjectivité des descriptions de documents ou d’artefacts.

4.1.4 Typologie, Espace, et Temps

Pour A. Gallay (cf. [GALL 86], l’activité archéologique se décompose en trois phases : la description (constructions « compilatoires »), la classification (constructions typologiques), l’interprétation (constructions explicatives). L’étape pivot, à savoir la classification, se fait selon les trois dimensions archéologiques identifiées par R. Etienne (cf. [ETIE 91]) : l’espace, le temps et les groupes (ou typologies).

En ce qui concerne la modélisation de ces classifications, la tentation consisterait à diviser ces dimensions hiérarchiquement, en prenant soin d’être complet et de ne pas se faire chevaucher les entités. Ceci a le mérite de simplifier la tâche, mais ne décrit évidemment pas la réalité archéologique. A contre pied, J. Litvak King (cf. [LITV 72]) suggère d’utiliser les diagrammes de Venn de la théorie des ensembles, apportant,

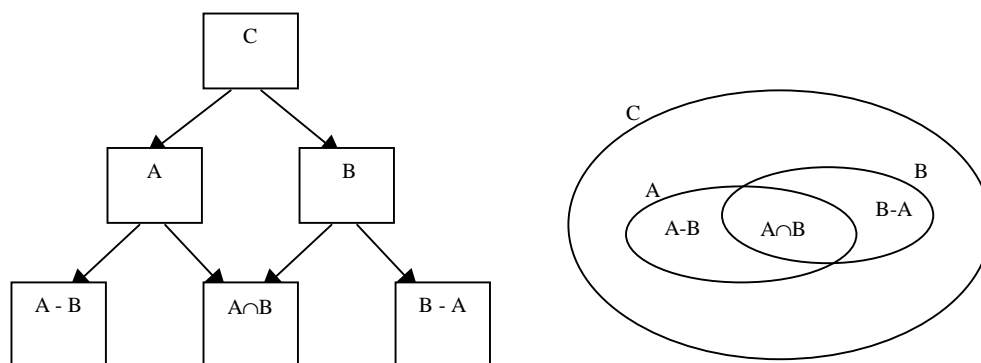


Figure 7 – Exemple de représentation d’inclusions et d’intersections d’ensembles sous forme d’un treillis.

entre autres, en plus de la relation d'inclusion, la relation d'intersection.

Ces relations d'inclusion et d'intersection peuvent être représentées par un treillis (cf. Figure 7) et donc *a fortiori* par notre modèle basé sur un ordre partiel. Notre système est donc tout à fait adapté à la formalisation des classifications temporelles, spatiales et typologiques.

4.2 Validation opératoire

A l'opposé de la validation par rapport aux besoins qui a fait l'objet d'une étude assez complète, la validation opératoire de la partie formelle du système n'est à l'heure actuelle qu'à l'état d'ébauche de ce qui pourra être réalisé dans le cadre d'une thèse. Un prototype en Prolog (cf. Annexe) a été implémenté en « traduisant » (de manière presque directe) l'axiomatique du système. Son application aux exemples donnés dans ce mémoire a conclu à une « non réfutation ». Nous envisageons pour le futur d'éprouver le système avec des jeux de tests aussi exhaustifs que possible.

Un autre point qui devra faire l'objet d'une attention soutenue lors du travail de thèse sera l'implantation du système sur une base de donnée professionnelle et l'étude de ses temps de réponse pour des volumes de données réalistes (de l'ordre de 10^6 descripteurs).

```
?- arcsConnus([1,2],L).
L = [universel,perses] ,
L = [perses,perses_gagnants] ,
no
?- arcsPossibles([1,2],L).
L = [universel,atheniens] ,
L = [universel,macedoniens] ,
L = [perses_gagnants,1] ,
L = [perses_gagnants,2] ,
no
?- arcsImpossibles([1,2],L).
L = [perses,perses_perdants] ,
no
```

Figure 8 – Exemple d'exécution du prototype (cf. Annexe).

5 Bilan et perspectives

L'idée que nous avons proposée est d'offrir aux experts un système leur permettant de prendre, retrouver et partager des « notes de lecture » dans un langage basé sur la composition formelle de descripteurs informels. 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... L'apport de l'outil informatique est de guider l'expert dans sa navigation entre les descripteurs en associant à cette navigation celle entre les corpus correspondants.

Le lecteur notera que nous limitons la mission de l'ordinateur à ce qui est purement *formel* et laissons à la charge de l'être humain ce qui concerne la *substance* des choses. Ceci est le résultat de l'application à l'indexation documentaire de la théorie « *situationnelle* » de la sémantique selon G. Mounin. Selon une telle théorie, le « sens » ne peut se situer que du côté de l'être humain et non du langage utilisé, c'est pourquoi nous avons choisi de placer l'interactivité au cœur de la recherche de documents.

Dans le cadre d'une thèse, nous souhaiterions valider notre modèle en l'appliquant à divers projets représentatifs du raisonnement archéologique : la consultation d'un corpus de chroniques de fouille, celle d'une bibliothèque spécialisée, celle d'une photothèque scientifique ou enfin celle de typologies d'artefacts. Ces applications auraient pour but d'approfondir tant les aspects opératoires que sémantiques du modèle. Nous projetons d'apporter une attention toute particulière à l'application de notre modèle pour les classifications spatiales, temporelles et typologiques, et à son utilisation comme support des interactions homme-machine.

6 Bibliographie

6.1 Informatique, Sciences de l'information

- [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.
- [BENE 99] **BENEL A., CALABRETTO S., PINON J.-M.** Indexation « sémantique » de documents archéologiques. A paraître dans : *Actes du deuxième colloque du chapitre français de l'ISKO, « L'indexation à l'heure d'Internet », Lyon, 21-22 Octobre 1999.* 7p.
- [BUSH 45] **BUSH V.** As We May Think. In: *The Atlantic Monthly*. July 1945.
- [CALA 93] **CALABRETTO S.** *Contribution à la validation des spécifications algébriques et à l'étude des spécifications algébriques avec contraintes.* Thèse de doctorat en informatique, Institut National des Sciences Appliquées de Lyon, 1993.
- [CHIA 99] **CHIARAMELLA Y.** Approches et modèles en recherche d'informations. In : *XVII^e congrès INFORSID. La Garde, France, 2-4 juin 1999.*
- [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.*
- [FOUR 98] **FOUREL F.** *Modélisation, indexation et recherche de documents structurés.* Thèse de doctorat en informatique, Université de Grenoble 1, 1998.
- [FUTR 94] **FUTRELLE, Robert P., ZHANG, Xiaolan.** Large - Scale Persistent Object Systems for Corpus Linguistics and Information Retrieval. In: *Digital Libraries '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station (Texas), June 19-21 1994.*

- [GARD 86] **GARDIN J.-C.** Systèmes experts et publications savantes. In: *The Fifth British Library Annual Research Lecture, 1986.*
- [GENE 99] **GENEST D.** Vers un système de recherche documentaire basé sur les graphes conceptuels. In: *Actes du XVII^e congrès INFORSID. La Garde, France, 2-4 juin 1999.* p.115-131.
- [LEHM 94] **LEHMANN P.** CCAT: The Current Status of the Conceptual Catalogue (Ontology) Group, with Proposals. In: *Proceedings of the Fourth International Workshop on Peirce: A conceptual Graph Workbench. Maryland, August 19, 1994.* p.18-28.
- [LESK 96] **LESK M.** *Seven Ages of Information Retrieval.* Occasional Paper. Universal Dataflow and Telecommunication Core Programme, International Federation of Library Associations and Institutions, 1996.
- [MOUN 97] **MOUNIN G.** *La sémantique.* Réédition revue et corrigée de l'édition de 1972. Editions Payot, 1997.
- [MECH 95] **MECHKOUR M., BERRUT C., CHIARAMELLA Y.** Using a Conceptual Graph Framework for Image Retrieval, In: *The International Conference on Multi-Media Modeling MMM'95, Nov. 14-17, 1995.*
- [OUNI 98] **OUNIS I., PASCA M.** *Modeling, Indexing and Retrieving Images using Conceptual Graphs.* MRIM Research Report, 1998.
- [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.*

6.2 Archéologie

- [COUR 82] **COURBIN P.** *Qu'est-ce que l'archéologie ? Essai sur la nature de la recherche archéologique.* Editions Payot, 1982.
- [DESF 91] **DEFARGES P., HELLY B.** L'archéologie, système d'information scientifique. In : *Aplicaciones Informaticás en Arqueologia : Teorias y sistemas. Saint-Germain-en-Laye, 1991.*
- [ETIE 91] **ETIENNE R., AUDA Y., IACOVELLA A.** Spécificité des problèmes d'analyse des données en archéologie : Application à l'analyse des nécropoles. In : *Aplicaciones Informaticás en Arqueologia : Teorias y sistemas. Saint-Germain-en-Laye, 1991.*
- [GALL 86] **GALLAY A.** *L'archéologie demain.* Belfond, 1986.
- [GARD 84] **GARDIN J.-C.** Les bases de données dans les sciences de l'antiquité : l'ajustement nécessaire des fins aux moyens. In : *Banques de données et sciences de l'antiquité, mars 1984.*
- [GARD 96b] **GARDIN J.-C.** Formalisation et simulation des raisonnements. In : *Une école pour les sciences sociales.* Ed. du Cerf, Ed. de l'Ecole des Hautes Etudes en Sciences Sociales. Paris, 1996. p.185-208.
- [GINO 78] **GINOUVES R., GUIMIER-SORBETS A.-M.** *La constitution des données en archéologie classique.* Ed. du CNRS, 1978.
- [LITV 72] **LITVAK KING J., GARCÍA MOLL R.** Set Theory Models: an Approach to taxonomic and locational relationships. In: *Models in Archaeology.* Methuen, London, 1972.

7 Annexe : Prototype en Prolog

% Regles de traitement des ensembles

membre(X,[X|_]).

membre(X,[_]Y):-
membre(X,Y).

nEstPasContenuDans(L,M):-
membre(X,L),not(membre(X,M)).

nEstPasDisjointDe([],_).

nEstPasDisjointDe(_,[]).

nEstPasDisjointDe(L,M):-
membre(X,L),membre(X,M).

estDisjointDe(L,M):-
not(nEstPasDisjointDe(L,M)).

contient(L,M):-
not(nEstPasContenuDans(M,L)).

% Regles de traitement des taxinomies (cf. specifications en 3.2)

descendantTerminal(X,X):-
fils(absurde,X).

descendantTerminal(X,Z):-
fils(Y,Z),descendantTerminal(X,Y).

aPourCorpus(X,L):-
setof(Y,descendantTerminal(Y,X),L).

rendConnu(L,X):-
aPourCorpus(X,M),contient(M,L).

rendPossible(L,X):-
aPourCorpus(X,M),nEstPasDisjointDe(L,M),
nEstPasContenuDans(L,M).

```

rendImpossible(L,X):-
    aPourCorpus(X,M),estDisjointDe(M,L).

arcsPossibles(L,[X,Y]):-
    rendConnu(L,X),fils(Y,X),rendPossible(L,Y).

arcsImpossibles(L,[X,Y]):-
    rendConnu(L,X),fils(Y,X),rendImpossible(L,Y).

arcsConnus(L,[X,Y]):-
    rendConnu(L,X),fils(Y,X),rendConnu(L,Y).

```

```

% Faits correspondant à l'exemple de la figure 2

```

```

fils(atheniens,universel).
fils(perses,universel).
fils(macedoniens,universel).
fils(atheniens_perdants,atheniens).
fils(atheniens_gagnants,atheniens).
fils(perses_perdants,perses).
fils(perses_gagnants,perses).
fils(macedoniens_perdants,macedoniens).
fils(macedoniens_gagnants,macedoniens).
fils(1,atheniens_perdants).
fils(1,perses_gagnants).
fils(3,atheniens_gagnants).
fils(3,perses_perdants).
fils(3,macedoniens_gagnants).
fils(2,macedoniens_perdants).
fils(2,perses_gagnants).
fils(absurde,1).
fils(absurde,2).
fils(absurde,3).

```