

« EXPLORATION » DE CORPUS DE DOCUMENTS ARCHEOLOGIQUES A L'AIDE DE THEORIES ALGEBRIQUES

AURELIEN BENEL, SYLVIE CALABRETTO *

Résumé

Cet article décrit un projet de collaboration avec l'Ecole Française d'Archéologie d'Athènes. L'objectif de ce projet est la mise en ligne de la « Chronique des Fouilles » publié au sein du « Bulletin de Correspondance Hellénique ». Cette chronique représente 80 ans d'Archéologie en Grèce et à Chypre dans de courts articles (environ 50000). Nous proposons un modèle d'indexation et une méthode de Recherche d'Information basés sur des théories algébriques. L'ensemble des descripteurs est structuré selon un ordre partiel afin de visualiser cette connaissance de manière graphique (graphe acyclique). La navigation au sein de ce graphe est assistée par un algorithme de filtrage basé sur la théorie des treillis. Ce système fournit aux archéologues un outil de prise de notes et de recherche de notes avec le minimum d'autorité extérieure afin d'assurer l'autonomie et la liberté nécessaire à leur fonction.

1. *Contexte de l'étude : la Chronique des fouilles de l'EFA*

L'Ecole française d'Athènes, établissement public de recherche dans les disciplines se rapportant au monde grec assure la diffusion de ses recherches grâce à ses publications. L'une des sources d'informations privilégiées pour les chercheurs en archéologie réside dans les chroniques de fouilles publiées au sein de l'organe de liaison de l'Ecole : le *Bulletin de Correspondance Hellénique*. Cette chronique a

* Aurelien.Benel@lisi.insa-lyon.fr, Sylvie.Calabretto@lisi.insa-lyon.fr

pour mission de signaler aux lecteurs toutes les "nouveautés" archéologiques en Grèce et à Chypre (fouilles, prospections, trouvailles fortuites, restaurations, muséologie, publications de matériel inédit) sur lesquelles des informations fiables ont été obtenues au cours de l'année. C'est aujourd'hui près de quatre-vingt ans d'archéologie qui y sont consignés. C'est dans ce contexte d'un corpus spécialisé adressé à des experts que s'inscrit notre travail visant à offrir des outils informatiques pouvant faciliter leurs tâches documentaires. Nous présenterons les principes mis en œuvre lors de la conception de notre système pour ensuite développer les bases algébriques de notre algorithme d'aide à l'exploration de corpus.

2. De la bibliothèque à l'Internet

2.1. Etudes expérimentales

Notons bien qu'il serait vain, dans des domaines nécessitant une expertise si haut niveau et une totale liberté, de prétendre établir un « modèle du domaine » à l'image de ce que l'on peut obtenir en informatique de gestion. Cependant, on peut tenter d'identifier, à l'aide d'études expérimentales, quelques activités documentaires typiques d'experts.

La première étude, réalisée par l'équipe de Kenton O'Hara, psychologue cognitif, concernait vingt-cinq doctorants de Cambridge (cf. [OHAR98]). Après examen de leurs pratiques dans une bibliothèque universitaire, il apparaît qu'en plus de la traditionnelle recherche bibliographique, les doctorants font des photocopies, lisent, relisent pour vérifier par exemple des références, rédigent, annotent en marge des photocopies ou prennent des notes à part : un ensemble de tâches qui montrent que l'utilisateur des bibliothèques ne fait pas que « consommer du savoir » mais qu'il en produit également.

Dans un autre domaine, les interviews menés par Andreas Paepcke portaient sur un grand nombre d'ingénieurs en entreprise et leur comportement en centre de documentation (cf. [PAEP96]). Ce qui ressort de l'étude est l'omniprésence des interactions sociales en ce qui concerne autant la découverte des informations que leur gestion, leur interprétation ou leur partage.

Les résultats de ces deux études sont d'autant plus intéressantes que les thèmes de l'*apport de l'utilisateur* et des *interactions sociales* ont été rarement pris en compte dans la conception des bibliothèques virtuelles (à l'exception de travaux tels que [GOH00], [ROSC95] et [TOCH94]).

2.2. Présentation de *Porphyre 2000*

Porphyre 2000 est un système client-serveur permettant de créer, retrouver et partager des annotations de document.

Premièrement, le serveur de document stocke les textes (RTF, XML, HTML) et images (GIF, JPEG) et permet d'en extraire l'intégralité ou des fragments.

Deuxièmement, le serveur de "stabylo" stocke les limites des parties de document surlignées par l'utilisateur.

Troisièmement, le serveur de graphe stocke et calcule les structures d'annotation liées à des fragments de documents ou des parties surlignées.

Du côté du client, différentes interfaces peuvent être utilisées suivant le rôle de l'utilisateur : expert ou responsable d'une communauté d'experts.

L'expert peut consulter les corpus mis à disposition par navigation dans les hypertextes (cf. Fig.1, fenêtres de droite) ou par navigation dans le graphe des annotations (cf. Fig.1, fenêtre de gauche). Chaque action de navigation réalisée dans l'une des deux fenêtres met à jour les deux.

L'expert peut surligner une partie du document et la lier à sa structure d'annotation. Enfin, il est en mesure de modifier sa structure personnelle d'annotation et de demander au responsable d'une communauté de la publier (au sens de « rendre public »).

Le responsable d'une communauté d'experts peut modifier les structures d'annotation soumises pour publication, les relier à d'autres et les rendre publiques (pour la communauté). Il est également en mesure d'ajouter de nouveaux documents dans le corpus de la communauté. Dans le cas de documents XML, le responsable dispose d'un « parser » intégré permettant de vérifier que le document obéit à notre définition de type de document (DTD).

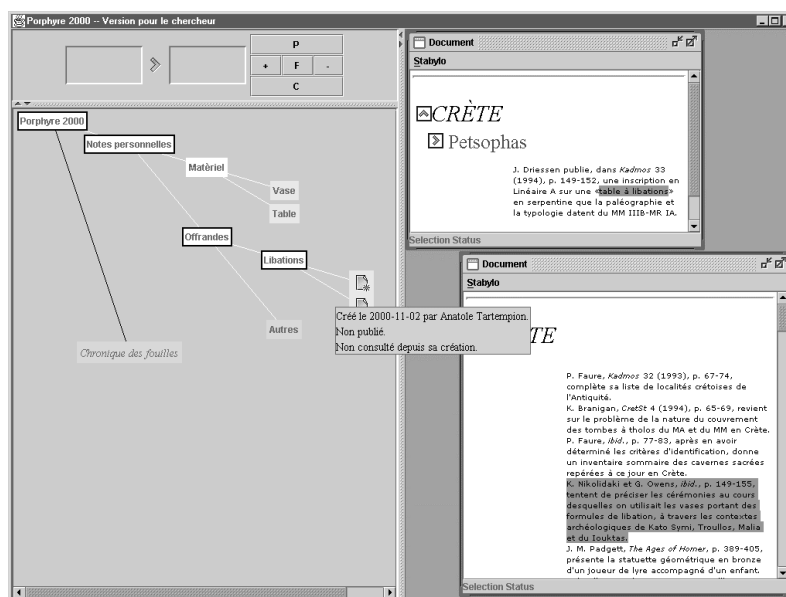


Figure 1 - Porphyre 2000 : Copie d'écran.

3. Algorithme de filtrage de graphe

La plupart des systèmes interactifs de recherche d'information (cf. [HEAR99]) se sont attachés à filtrer les informations afin d'en réduire la charge cognitive.

Dans notre système, étant donné la taille importante que prendra le graphe d'annotations au fur et à mesure de son utilisation, il est crucial de ne présenter à un moment donné qu'une partie de ce graphe. C'est cette partie, fortement basée sur des théories algébriques que nous allons maintenant présenter en détail.

3.1. L'approche booléenne revue et corrigée

Gerard Salton à la fin des années 60 (cf. [SALT68]) a défini un modèle pour la recherche d'information basé sur la théorie des ensembles. Ce modèle considère un ensemble des documents et un ensemble des « descripteurs ». Ainsi on peut tracer les graphes d'inclusion de corpus (cf. Fig.2) et de conjonction de requêtes (cf. Fig.3).

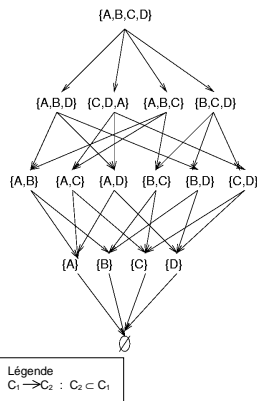


Figure 2 - Structure en treillis de l'espace des documents.

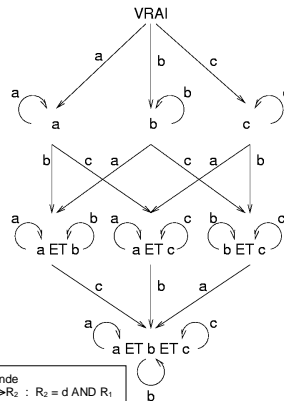


Figure 3 - Structure en treillis de l'espace des descripteurs.

Il devient ensuite possible de déduire de la correspondance entre documents et descripteurs (cf. Tab.1), la correspondance entre requêtes et corpus (cf. Tab.2). De là, on remarque que certains corpus ne peuvent être obtenus par aucune requête (ex : $\{B,C\}$) et que le même corpus peut être obtenu par différentes requêtes (ex : la requête a ET b et la requête b).

Tableau 1- Exemple de correspondance entre les documents A, B, C, D et les descripteurs a, b, c.

		Descripteurs		
		a	b	c
Documents	A	X		X
	B	X	X	
	C			X
	D	X		

Tableau 2 - Exemple de correspondance entre les requêtes et les corpus de documents (calculé à partir du tableau 1).

VRAI	$\{A,B,C,D\}$
a	$\{A,B,D\}$
b	$\{B\}$
c	$\{A,C\}$
a ET b	$\{B\}$
b ET c	\emptyset
c ET a	$\{A\}$
a ET b ET c	\emptyset

Ces résultats tout aussi connus qu'ils soient, ont été fort peu utilisés comme support des interactions homme-machines. Claudio Carpineto (cf. [CARP94]) les a utilisés en enlevant du graphe d'inclusion des corpus les corpus inaccessibles, obtenant ainsi un diagramme statique de généralisation/spécialisation des classes de document (cf. Fig.4).

Dans notre approche (cf. nos précédents travaux [BENE99], [BENE00a] et [BENE00b]), nous fusionnons dans le graphe des requêtes, celles qui décrivent le même corpus. Nous obtenons ainsi un diagramme d'état (cf. Fig.5) dans lequel les états correspondent à des corpus et les transitions à des requêtes élémentaires. Ces requêtes à un seul descripteur sur des corpus intermédiaires peuvent être vues comme l'ajout d'un descripteur à la requête globale : il s'agit d'une manière de « raffiner » la requête (en anglais : « query refining »).

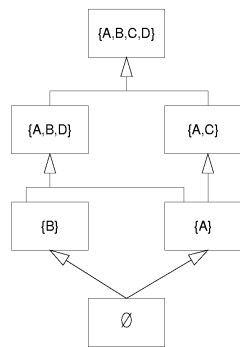


Figure 4 - Diagramme de classe (cf. [UML97]) dérivé de l'espace des documents.

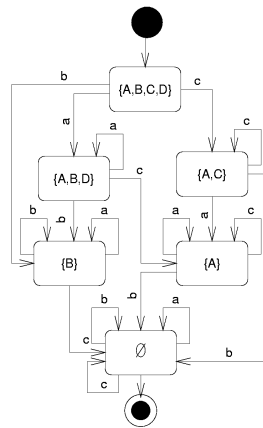


Figure 5 - Diagramme d'état (cf. [UML97]) dérivé de l'espace des descripteurs.

3.2. Le diagramme d'état comme filtre de graphe

Si nous reprenons le diagramme d'état précédent (cf. Fig.5), dans un corpus donné chaque descripteur peut être dit :

- *impossible* : s'il mène du corpus actuel au corpus vide (ex : le descripteur *c* dans l'état $\{B\}$),
- *possible* : s'il « boucle » sur le corpus actuel (ex : le descripteur *a* dans l'état $\{B\}$),
- *possible* : dans les autres cas.

Nous proposons d'utiliser ces définitions comme un filtre dynamique sur un graphe de descripteurs. Ce graphe, orienté et acyclique, représente un ordre partiel signifiant que pour deux descripteurs D_1 et D_2 , $D_1 > D_2$ si et seulement si tout descripteur décrit par D_2 l'est aussi par D_1 .

Le filtre consiste, comme nous allons le voir dans l'exemple de la section suivante, à montrer uniquement les descripteurs *connus* ainsi que leurs inférieurs directs et à assigner à chaque descripteur son état (possible, impossible, connu).

3.3. Scénario de recherche de documents

Suivons pas à pas le scénario présenté dans la figure 6 :

Étape 1. Le corpus global traite de « *vestiges typés* ». Les corpus plus spécialisés traitent de *vestiges datés*, de *vestiges de type épigraphique* ou de *vestiges de type instrumenta/mobilier* mais pas de *vestiges de type architectural* (cette description ne correspond en effet à aucun document du corpus considéré). Après sélection par l'utilisateur de *instrumenta/mobilier*, le système passe à l'étape 2.

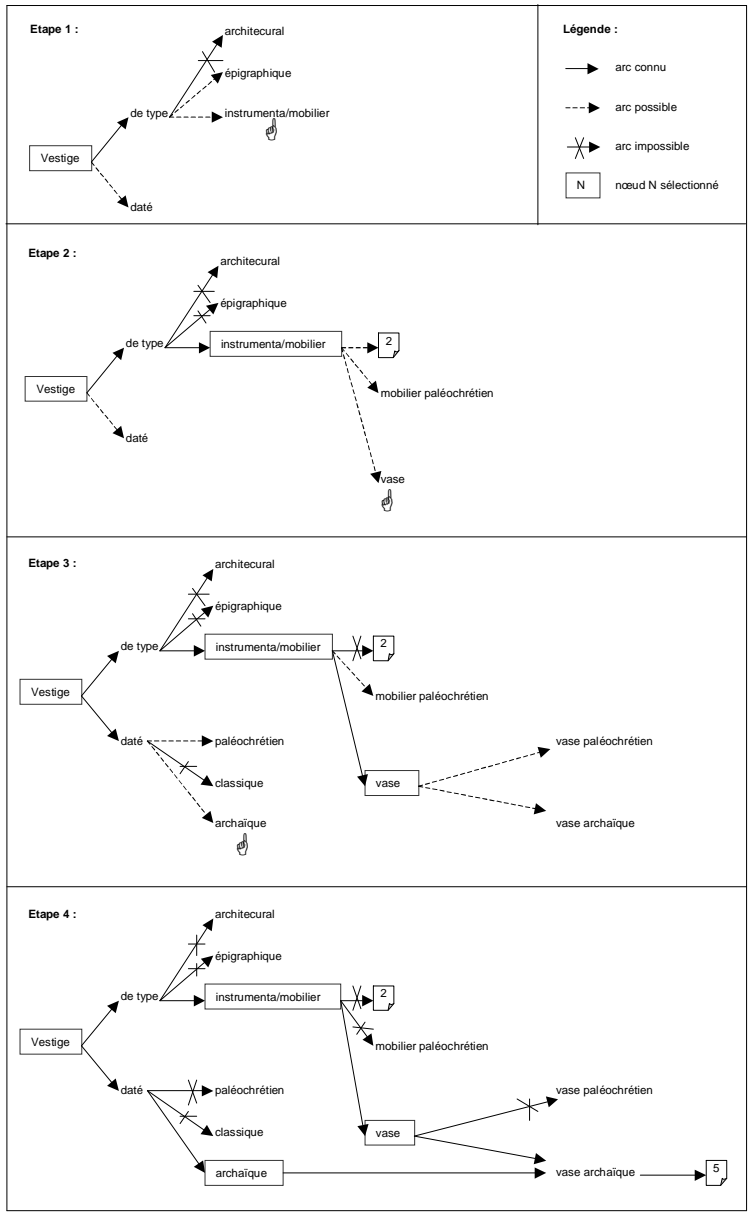


Figure 6 – Scénario de navigation dans un graphe d’annotations.

Etape 2. Le corpus sélectionné traite de *vestiges de type instrumenta/mobilier*. Cette description correspond exactement au document ayant 2 pour identifiant. Les corpus plus spécialisés traitent de *mobilier paléochrétien* ou de *vases*. Aucun ne traite de *vestiges de type architectural* ou *épigraphique*. Après sélection par l'utilisateur de *vase*, le système passe à l'étape 3.

Etape 3. Le corpus sélectionné traite de *vestiges de type vase (instrumenta/mobilier)* et de *vestiges datés*. On remarque que le fait qu'ils soient *datés* est inféré par l'ordinateur (tous les documents du corpus traitant de *vases* traitent de *vestiges datés*). Cette description correspond exactement au document ayant 2 pour identifiant. Les corpus plus spécialisés traitent de *mobilier paléochrétien*, de *vases paléochrétiens*, de *vases archaïques*, de *vestiges paléochrétiens* ou de *vestiges archaïques*. Aucun ne traite de *vestiges de type architectural* ou *épigraphique*, ni de *vestiges datés de l'époque classique*. Après sélection par l'utilisateur d'*archaïque*, le système passe à l'étape 4.

Etape 4. Le corpus sélectionné traite de *vestiges de type vase (instrumenta/mobilier) datés de l'époque archaïque*. Ce corpus ne contient qu'un seul document : celui ayant 5 comme identifiant.

4. Bilan et perspectives

Nous nous sommes donc intéressés aux tâches documentaires d'experts (dans notre cas, des chercheurs en archéologie) face à un corpus spécialisé. Tout d'abord, nous avons intégré la recherche de documents dans une activité plus large, plus complexe, impliquant d'avantage l'utilisateur et comprenant entre autres l'enrichissement de la description des documents et le partage de cette description avec d'autres utilisateurs. Ensuite nous avons introduit une approche de la recherche de

documents basée sur la navigation entre corpus, mettant ainsi en avant l'utilisateur, le corpus de documents et leurs relations (interactivité, aspect exploratoire...). Enfin nous avons défini comment un outil filtrant un graphe d'indexation en fonction de la sélection de descripteurs permettait de rechercher et d'indexer des documents ainsi que de les annoter et d'échanger ces annotations.

Nous prévoyons d'éprouver d'ici peu notre prototype, tant au niveau algorithmique avec de gros volumes de données, qu'au niveau interface auprès des utilisateurs.

Ainsi, dans les prochaines années, nous pourrons offrir à des communautés d'experts un système leur permettant de prendre des notes de lecture, de les partager, et de *naviguer* à travers le réseau qu'elles constituent ; le pari étant, tel que le souhaitait Vannevar Bush (cf. [BUSH45]), d'assister le lecteur 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.

REFERENCES

- [BENE00a] Bénel, A., Calabretto, S., Pinon, J.-M., Iacovella, A. Vers un outil documentaire unifié pour les chercheurs en archéologie. In : *Actes du XVIIIe congrès INFORSID, 2000*. pp.133-145.
- [BENE00b] Bénel, A., Calabretto, S., Pinon, J.-M., Iacovella, A. Consultation de documents et sémantique : Application à des publications savantes. In : *Actes du second Colloque International Francophone sur l' Ecrit et le Document, 2000*. pp.271-280.
- [BENE99] Bénel, A., Calabretto, S., Pinon, J.-M. Indexation "sémantique" de documents archéologiques. A paraître in : *Actes du deuxième colloque du chapitre français de l' SKO, "L' idexation à l'heure d' Internet", 1999*.
- [BUSH45] Bush, V. As we may think. In : *The Atlantic Monthly. July 1945*.

- [CARP94] Carpineto, C., and Romano, G. Dynamically bounding browsable retrieval spaces: an application to Galois lattices. In : *RIA'O'94 conference proceedings, "Intelligent Multimedia Information Retrieval Systems and Management"*, 1994.
- [GOH00] Goh, D., Leggett, J. Patron-augmented digital libraries. In : *Proceedings of the Fifth ACM Conference on Digital Libraries, 2000*.
- [OHAR98] O'Hara, K., Smith, F., Newman, W., and Sellen, A. Student readers' use of library documents: implications for library technologies. In : *ACM Conference Proceedings on Human Factors in Computing Systems, 1998*.
- [HEAR99] Hearst, M. User interfaces and visualization. In : Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*, ACM Press and Addison Wesley, 1999.
- [PAEP96] Paepcke, A. Digital libraries: Searching is not Enough. What we learned on-site. In : *D-Lib Magazine. May 1996*.
- [ROSC95] Röscheisen, M., Mogensen, C., and Winograd, T. Beyond browsing: shared comments, soaps, trails and on-line communities. In : *the Third International World Wide Web Conference, "Technology, Tools and Applications"*, 1995.
- [SALT68] Salton, G. *Automatic Information Organization and Retrieval*. Chapter: "Retrieval models". Computer Sciences series, McGraw-Hill Inc., 1968.
- [SOWA92] Sowa J.F. Semantic networks. In : Shapiro, S.C. *Encyclopedia of Artificial Intelligence*, Wiley, New York, 1992.
- [TOCH94] Tochtermann, K. A first step toward communication in virtual libraries. In : *the Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, 1994*.
- [UML97] *UML Notation Guide*. OMG, 1997.