



Ecole Française d'Athènes  
Service Informatique

# LA CHRONIQUE DES FOUILLES DE LA BIBLIOTHEQUE A L'INTERNET



# LA CHRONIQUE DES FOUILLES : DE LA BIBLIOTHEQUE A L'INTERNET

## Etude globale du projet

Title	Global Study of the Project
Author	Aurélien BÉNEL ( <a href="mailto:aurelien_benel@writeme.com">aurelien_benel@writeme.com</a> )
Project	The Excavations Chronicle : from Library to Internet
Organisation	French School of Archaeology (Athens)
Created on	June 11, 1998
Modified on	October 27, 1998
Version	Final draft
Language	French – standard
Keywords	Virtual library, electronic documents, information retrieval, scanning, OCR, semantic model, metadata.

## INTRODUCTION

Le présent document a pour mission de fournir une trace écrite du stage effectué du 15 mai au 15 septembre 1998 au sein du service informatique de l'Ecole française d'Athènes, sous la direction de Monsieur Andréa Iacovella. Son objectif est double, il s'agit d'une part de rendre compte des points mis en lumière par l'étude et des recommandations qui en découlent, mais également de décrire le processus qui a rendu cette étude possible. Dans une première partie, nous décrirons donc les méthodes et l'organisation mises en œuvre durant le déroulement du stage, et dans une deuxième partie, les orientations scientifiques découlant du projet.

<b>LA CHRONIQUE DES FOUILLES : DE LA BIBLIOTHÈQUE À L'INTERNET.....</b>	<b>2</b>
ETUDE GLOBALE DU PROJET .....	2
<b>INTRODUCTION .....</b>	<b>2</b>
<b>I. MÉTHODE ET ORGANISATION .....</b>	<b>4</b>
A. IDENTIFICATION DU PROJET .....	4
1. <i>Problématique générale</i> .....	4
2. <i>Finalité</i> .....	4
3. <i>Description</i> .....	4
B. DÉROULEMENT DU STAGE .....	5
1. <i>Nature des tâches</i> .....	5
2. <i>Evolution de l'appréhension du problème</i> .....	5
<b>II. RAPPORT SCIENTIFIQUE.....</b>	<b>7</b>
A. L'HÉRITAGE DE LA CHRONIQUE DES FOUILLES.....	7
1. <i>Objectifs et histoire de la publication</i> .....	7
2. <i>Estimation Volumétrique</i> .....	7
3. <i>Richesse et complexité</i> .....	7
B. NUMÉRISATION ET PERTES D'INFORMATIONS .....	9
1. <i>Echantillonnage</i> .....	9
2. <i>Filtrage</i> .....	9
3. <i>Un "monde de blocs"</i> .....	10
C. UN NOUVEAU SUPPORT POUR L'ÉCRITURE.....	12
1. <i>Formats : le cauchemar de la multitude et de la normalisation</i> .....	12
2. <i>Signifiant/Signifié : Que stocker ?</i> .....	13
3. <i>Quels choix pour quels besoins ?</i> .....	16
D. RECHERCHE D'INFORMATIONS TEXTUELLES .....	18
1. <i>Recherche intellectuelle</i> .....	18
2. <i>Recherche statistique</i> .....	19
E. VERS UNE SOLUTION .....	22
1. <i>Pourquoi des documents électroniques ?</i> .....	22
2. <i>Et les périodiques scientifiques ?</i> .....	22
3. <i>Que stocker ?</i> .....	22
4. <i>Sous quelle forme ?</i> .....	22
5. <i>Du papier au numérique, quelle voie ?</i> .....	22
6. <i>Peut-on parler de bibliothèque virtuelle ?</i> .....	23
<b>CONCLUSION.....</b>	<b>24</b>

# I. METHODE ET ORGANISATION

## A. IDENTIFICATION DU PROJET

### 1. Problématique générale

L'Ecole française d'Athènes, établissement public de recherche dans les disciplines se rapportant au monde grec, est dotée d'une riche collection documentaire (bibliothèque, photothèque, "planothèque", archives manuscrites) destinée à l'origine à ses membres recrutés sur concours. Depuis quelques années, l'Ecole accueille des universitaires et des chercheurs du monde entier et leur fait profiter de cet inégalable outil de travail. Dans le but de faire rayonner ce dernier au-delà des murs de l'Ecole, un vaste programme a été entrepris afin de mettre à disposition sur Internet des services documentaires destinés aux chercheurs. Ces services consisteront, suivant la nature des documents, à l'accès à des inventaires informatisés ou bien aux documents eux-mêmes sous une forme électronique.

### 2. Finalité

L'une des sources d'informations privilégiées pour les chercheurs en archéologie réside dans les chroniques de fouilles. L'une des plus célèbres est celle de l'Ecole française d'Athènes, publiée au sein de son organe de liaison : le *Bulletin de Correspondance Hellénique*. Cette chronique des fouilles, a pour mission de signaler aux lecteurs toutes les "nouveauétés" 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 80 ans d'archéologie qui y sont consignés. Cette incomparable source d'information pâtit cependant des limitations inhérentes à la nature physique des ouvrages "papier" : le nombre d'exemplaires est limité (surtout pour les numéros les plus anciens), le transport des ouvrages ou de fac-similés est coûteux et délicat, enfin la recherche d'une information spécifique est une tâche fort longue et ingrate. Les progrès de l'informatique nous permettent de considérer d'autres solutions plus adaptées. En effet, pour peu qu'un document existe sous une forme électronique, il peut, via un réseau mondial d'informations comme Internet, être mis à la disposition d'un très grand nombre de lecteurs et ceci par delà les mers et les frontières, de plus ces lecteurs peuvent profiter de la puissance de traitement des ordinateurs pour rechercher des informations.

### 3. Description

Il s'agit d'offrir à la communauté scientifique un service, adapté aux besoins de la recherche en archéologie, lui permettant d'accéder à distance aux connaissances contenues dans la chronique des fouilles du *Bulletin de Correspondance Hellénique*. Ceci comprend la conversion sous une forme numérique des publications existant sous forme papier, l'intégration du livrable numérique dans la chaîne de production des publications à venir, la mise en place de services, humains et automatiques, de recherche et de consultation via Internet.

L'étude globale du projet doit permettre de définir quels domaines des sciences de l'information et des disciplines connexes sont requises par un tel projet, quels en sont les technologies et les théories directement applicables, ou qui sont encore du domaine de la recherche.

## **B. DEROULEMENT DU STAGE**

### **1. Nature des tâches**

La nature des tâches effectuées durant le stage fut de plusieurs ordres : analyse du corpus à traiter, tests d'outils, collecte d'informations sur Internet, correspondance (en anglais) avec des chercheurs et des industriels, rédaction de textes de synthèse.

Pour ce qui est de la consultation du corpus, le but était de me familiariser avec la *materia prima* du projet, et de mettre en évidence ses particularités quantitatives et qualitatives qui auraient une conséquence sur le déroulement futur du projet. Il est à noter que l'évaluation quantitative exhaustive (nombre de pages, pourcentage respectif d'occupation par les images et par le texte) avait un caractère des plus ingrats, et aurait mérité, une fois que j'aurais posé les directives, d'être effectuées par plusieurs personnes (et pas forcément des informaticiens !). Pour que l'étude soit complète, il reste d'ailleurs à recenser le nombre de figures de la chronique.

Les tests d'outils logiciels et matériels portèrent principalement sur l'influence des paramètres de numérisation et de compression sur les résultats de reconnaissance optique de caractères et sur la qualité de la visualisation. Face au nombre de combinaisons possibles des différents paramètres, j'ai évolué quelque peu à tâtons. On peut s'interroger maintenant sur la rentabilité de cette tâche, connaissant les résultats obtenus et le temps consacré.

En ce qui concerne la collecte d'informations sur la *World Wide Web*, il est intéressant d'en toucher du doigt les limites. La recherche d'informations, tout d'abord, en est réduite à la recherche d'occurrences de suites de caractères, il est impossible de spécifier le type de documents que l'on souhaite (articles scientifiques, publicités, document de vulgarisation...), la recherche s'étend à toutes les pages du WWW et non aux quelques sites spécialisés, la formulation d'une question reste très difficile pour quelqu'un qui ne connaît pas parfaitement le domaine et les termes « consacrés ». Enfin, les documents eux-mêmes répondent mal aux interrogations qu'un tel projet peut susciter. Qu'il s'agisse de comptes-rendus de colloques, ou de publicités, les reproches qu'on peut leur faire et le peu de distance prise par rapport au cas étudié, aux technologies utilisées, et aux partenaires industriels impliqués. Il manque cruellement d'études synthétiques, lisibles par des non-techniciens, et dont la validité dépasserait le caractère éphémère du marché.

Au sujet de ma correspondance, il est nécessaire de noter qu'elle fut précédée par ma recherche sur le WWW. C'est là en effet, que j'ai pu d'une part apprendre les bases du domaine et d'autre part apprendre les coordonnées des entreprises spécialisées (travaux de numérisation, de saisie, éditeurs de logiciels d'OCR, de moteurs de recherche...) et des listes de discussion. L'apport des sociétés commerciales ne fut que d'une importance relative, étant donnée leur attitude partielle, et la restriction de leurs connaissances au domaine strictement technologique. Par contre, le fait de poser une question à une liste de discussion m'a permis, par les réponses que j'ai reçu, d'édifier un réseau de correspondants (chercheurs, bibliothécaires, professeurs) impliqués dans des projets comparables et prêts à m'aiguiller, au fil de mes questions, vers un article particulier, une technologie donnée, ou un point à approfondir.

### **2. Evolution de l'appréhension du problème**

Devant le grand nombre de facettes du problème, j'ai été amené au cours du stage à en étudier en détail l'une ou l'autre, à en découvrir d'autres insoupçonnées, à prendre du recul par rapport aux séparations habituelles entre domaines de recherche, à tenter de faire des synthèses. C'est ainsi, que partant de l'idée qu'il s'agissait d'un problème de numérisation et de reconnaissance optique de caractères, j'en suis venu tour à tour à le poser comme étant un sujet se rattachant aux bibliothèques virtuelles, à la recherche d'informations, aux documents électroniques, pour finalement aboutir à la modélisation sémantique de documents. Il faut souligner que tous les virages importants de mon étude (distinction entre l'aspect et le contenu d'un document, modèles sémantiques, langages de linéarisation) ont été initiés par ma correspondance avec des spécialistes du domaine. Même s'ils n'ont jamais posé ces problèmes dans ces termes, c'est en citant un article, une technologie particulière qu'ils m'ont mis sur la voie. Ce fait, à considérer avec beaucoup d'intérêt pour les méthodes à employer dans les projets futurs, nous éclaire également sur le projet lui-même, à savoir la complémentarité pour l'apprentissage dans une bibliothèque (virtuelle ou non) entre les documents qui apportent les connaissances du domaine et les interactions interpersonnelles qui donnent les grandes orientations. Pour finir, je voudrais indiquer que, souvent, à l'issue de mes « découvertes », me revenaient, *a posteriori*, des conversations, entretenues avec mon directeur de stage, qui, avait déjà une idée synthétique du problème. Je regrette un peu de ne pas avoir su relever

immédiatement les allusions qui auraient pu me guider dans les dédales de la bibliographie. Mais, probablement que mes errances m'ont été aussi bénéfiques pour ma formation que mes « découvertes »...

## II. RAPPORT SCIENTIFIQUE

### A. L'HERITAGE DE LA CHRONIQUE DES FOUILLES

Nous nous proposons ici d'analyser l'héritage que constitue le corpus de la *Chronique des fouilles* de l'Ecole française d'Athènes existant sur papier, dans l'optique d'une éventuelle transformation sous forme électronique. Dans un premier temps, nous identifierons le domaine d'étude que s'est fixé la chronique, et sa possible évolution au cours de son histoire. Dans un deuxième temps, nous estimerons le volume du corpus afin de déterminer l'ampleur d'un tel projet. Enfin, dans un troisième temps, nous tenterons de présenter les caractéristiques de la chronique qui rendent sa numérisation à la fois complexe et digne d'intérêt.

#### 1. Objectifs et histoire de la publication

Le *Bulletin de correspondance hellénique* [BCH] est l'organe de liaison des correspondants de l'Ecole française d'Athènes. Sa publication est annuelle et s'effectue en deux livraisons (voire trois ces dernières années). Le premier tome est traditionnellement consacré aux articles de synthèse, le second l'est à l'information sur les activités de l'Ecole et à la publication du matériel archéologique. Son premier numéro remonte à 1877, où il venait prendre la relève du *Bulletin de l'Ecole française d'Athènes* (douze livraisons entre 1868 et 1871).

Au sein de ce bulletin, la *Chronique des fouilles* a pour mission de signaler aux lecteurs toutes les "nouveauautés" archéologiques (fouilles, prospections, trouvailles fortuites, restaurations, muséologie, publications de matériel inédit) sur lesquelles des informations fiables ont été obtenues au cours de l'année. La chronique fait son apparition dans le bulletin de 1920. Tout d'abord appelée *Chronique des fouilles dans l'Orient hellénique*, sa portée géographique embrassait les limites de la Grèce antique. Elle adopte son titre actuel de *Chronique des fouilles en Grèce* en 1936, et se cantonne depuis aux frontières actuelles de la Grèce. Enfin en 1959, parallèlement à la chronique en Grèce, apparaît la *Chronique des fouilles à Chypre*.

Couvrant indifféremment, à l'origine, les fouilles de l'Ecole et les autres, la chronique se scinde, en 1940, en deux rubriques traitant respectivement des unes et des autres. Enfin, à partir de 1970, les travaux de l'Ecole n'apparaissent plus dans la chronique que sous forme de références à une section autonome du bulletin créée pour l'occasion.

#### 2. Estimation Volumétrique

Les volumes indiqués ici s'appuient sur l'inventaire réalisé en mai 1998 sur les 108 chroniques de fouilles disponibles, à savoir les 70 numéros de la *Chronique de fouille en Grèce* (et en Orient Hellénique) de 1920 à 1995 (sachant qu'il n'en existe pas pour 1932 et 1946, et que sont regroupés respectivement ceux de 1940 et 1941, 1942 et 1943, 1947 et 1948) et les 38 numéros de la *Chronique des fouilles à Chypre* de 1959 à 1996.

NOMBRE DE PAGES	OCCUPATION MOYENNE DES PAGES PAR LES FIGURES	NOMBRE DE CARACTERES	NOMBRE DE FIGURES
12.000 ± 500	(55,0 ± 2,5) %	30.000.000 ± 500.000	??? ± ???

#### 3. Richesse et complexité


##### a. Sens

Par la valeur sémantique des mots contenus dans la chronique, nous pouvons remarquer le grand effort de formalisation dont elle est le fruit. Elle est constituée de passages complètement indépendants les uns des autres, dont chacun traite en général d'une nouveauté archéologique sur un site géographique (la nouveauté pouvant être constituée de plusieurs vestiges), et pour lequel les types d'informations apportées sont souvent récurrents : lieu précis de la découverte, inventeur, nature des vestiges, datation, etc. La langue utilisée est le français, mais certains mots sont empruntés au grec (ancien et moderne), latin, italien, allemand, anglais...

## **b. Structure**

Le lien entre ces flots d'information est rendu par la structure de la chronique. On dénote une hiérarchie des titres particulièrement profonde (au moins sur cinq niveaux), qui permet de regrouper les passages tout d'abord par rapport à la géographie des sites dont ils traitent (donnant une impression de "zoom" d'un niveau à l'autre), mais aussi par rapport à la nature des nouveautés (muséologie, découvertes, etc.) ou encore par rapport à l'organisme effectuant les travaux archéologiques (Ecole française ou autre institution). Outre les titres, il est fait usage massivement d'autres liens entre les informations comme, par exemple, les appels de notes et de figures. Enfin, il est fait référence à des informations externes dans le cas des renvois bibliographiques.

## **c. Aspect**

La richesse que nous venons d'évoquer au sujet du sens et de la structure de la publication a pour conséquence une richesse comparable de la mise en page et de la typographie. On peut citer en premier la diversité des fontes d'imprimerie utilisées, à savoir : des caractères latins accentués, des caractères grecs avec accents et "esprits" (ex. : εἴνεκα), et enfin des fontes sans doute faites sur mesure restituant à l'identique une inscription, un monogramme, une assise de pierre, un crampon... (ex. : ΕΡΑΚΛΕΙ ). Il est important ensuite de préciser que les conventions associant une information structurelle à l'aspect de la page ont évolué avec l'histoire de la publication, et particulièrement dans la Chronique des fouilles à Chypre. C'est ainsi que les notes peuvent migrer du bas des pages à la fin du texte, les appels de figure peuvent se déplacer du corps du texte vers la marge, les titres passer de caractères gras à des caractères italiques, et ainsi de suite. Enfin, nous devons distinguer dans le matériel graphique de la chronique, différentes classes en fonction de la nature (photographies, dessins au trait, dessins en dégradés) et du mode d'insertion (figures, planches, hors textes).

Tout d'abord, par son histoire et son volume d'informations, la Chronique des fouilles représente un outil pratiquement unique pour les chercheurs : quatre-vingt ans d'activité archéologique en Grèce et trente ans à Chypre. Ensuite, malgré cette quasi exhaustivité, le volume du corpus reste suffisamment limité pour envisager sa numérisation dans des délais et des budgets raisonnables. Enfin, sa richesse de structure et d'aspect fait de sa conversion sous une forme numérique, un réel défi informatique, et la rend digne d'intérêt pour la communauté scientifique en informatique.

Dans le cadre du projet de création d'un corpus électronique à partir des publications sur papier de la *Chronique des fouilles*, nous ne disposons plus que d'une information visuelle. Une telle information portant sur l'apparence du document est très dégradée : elle nécessite en effet une interprétation pour retrouver son sens et est fortement récurrente (ce qui explique la quantité d'informations nécessaire pour la modéliser). Deux marches à suivre sont possibles : soit passer par l'interprétation d'un être humain et créer de toute pièce une information très signifiante (dessins vectoriels, textes saisis, etc.), soit capturer optiquement le matériel visuel et le faire passer ensuite par une série de filtres et de traitements afin d'en tirer la "substantifique moelle". Nous allons ici étudier la deuxième méthode qui présente l'intérêt de pouvoir être partiellement automatisée. Dans un premier temps, nous nous intéresserons au thème de la perte d'informations inhérente à toute acquisition numérique. Ensuite, nous mettrons en évidence, pour la *Chronique des fouilles*, les besoins en informations dus au matériel à numériser et à l'utilisation que l'on souhaite en faire. Enfin, nous définirons des recommandations pour les paramètres de numérisation à employer pour le projet.

## **B. NUMERISATION ET PERTES D'INFORMATIONS**

La numérisation consiste à modéliser une partie du monde réel par une suite de nombres. Ce processus se décompose en général en une première étape que nous appellerons "échantillonnage" et une seconde que nous nommerons "filtrage". Nous verrons que chacune d'elle s'accompagne de pertes d'informations par rapport à l'original.

### **1. Echantillonnage**

Dans cette phase, il va s'agir de réaliser des mesures sur le monde réel. Ceci ne peut se faire qu'avec des pertes d'informations étant donnée l'opposition entre le caractère continu et infini du monde réel d'une part et les restrictions apportées par les représentations numériques d'autre part.

#### **a. Incrément de numérisation**

Une infinité de mesures infiniment proches les unes des autres pourraient être réalisées sur le monde réel, cependant la numérisation entraîne la restriction de l'information à une suite finie de valeurs. Il est indispensable de comprendre qu'aussi fine serait la prise de mesure, elle serait incapable de rendre compte de l'infinité de données contenues dans le monde réel. Il faudra donc définir la taille de la donnée atomique en fonction de la précision nécessitée par l'information elle-même<sup>1</sup> ou son utilisation (traitement, consultation, etc.). Dans le cas d'une information visuelle, la fréquence (spatiale) d'échantillonnage est appelée résolution, traditionnellement exprimée en nombre de points par pouce<sup>2</sup>.

#### **b. Nombre de niveaux**

La deuxième perte concerne les "niveaux" : dans le monde réel, la mesure d'une donnée atomique prend ses valeurs parmi un ensemble infini (par exemple les nombres réels), la numérisation nécessitera une approximation des niveaux du monde réel à des niveaux appartenant à un ensemble fini de valeurs. Dans le cas d'une image, il s'agit des couleurs. Le nombre de couleurs utilisé est en général de 2 pour représenter les documents en noir et blanc, de 256 pour ceux en dégradés de gris, et de 16.777.216 pour ceux en couleur. Rappelons tout de même que tout dépend de l'utilisation que l'on veut en faire : par exemple la numérisation en noir et blanc d'une lettre manuscrite permettra tout à fait sa lecture mais non l'étude de la couleur du papier ou de l'encre.

#### **c. "Ré – échantillonnage"**

Notons pour finir qu'il est toujours possible par traitement de modifier *a posteriori* les paramètres établis lors de l'acquisition (incrément et nombre de niveaux), mais de tels traitements sont destructeurs d'informations. Dans le cas d'une réduction du nombre de données cela est évident, mais même dans le cas d'ajout de données par déduction, ceci peut être considéré comme une perte d'informations en ce sens qu'il devient impossible de distinguer les données réelles des données calculées. Ainsi par exemple, l'agrandissement par "ré - échantillonnage" d'un document visuel se traduit par un effet de flou.

### **2. Filtrage**

Cette phase consiste à transformer l'information numérique obtenue lors de la phase d'acquisition en une autre information numérique plus signifiante. La perte d'informations est ici intentionnelle, il s'agit de faire disparaître des informations risquant, selon le vieil adage des informaticiens, de "tuer l'information". Le problème dans n'importe quel filtrage est de définir le seuil entre l'information à détruire et celle à garder.

---

<sup>1</sup> Par exemple, dans le cas où une information est récurrente suivant une fréquence fixe (spatiale ou temporelle), un échantillonnage dont la fréquence serait moins de deux fois celle de l'information aurait pour conséquence l'apparition de parasites (effet moiré pour une image, bruit sourd pour un son). Si l'on souhaite conserver cette récurrence, on doit échantillonner à fréquence élevée, sinon on peut la supprimer avant échantillonnage par l'usage d'un filtre.

<sup>2</sup> En anglais : *dots per inch [dpi]*

### **a. "Détachage"**

Le détachage consiste à supprimer des motifs inutiles à l'objectif que l'on s'est fixé. Il est possible de le faire automatiquement, en détectant et supprimant les groupes de points isolés d'une taille inférieure à un seuil donné. La valeur du seuil est à fixer avec beaucoup d'attention pour éviter de supprimer des informations nécessaires. Ainsi, par exemple, il faut s'assurer lorsque l'on applique un tel traitement à l'image d'une page dactylographiée de vérifier que le point sur les "i" ne seront pas ôtés.

### **b. "Dilution"**

Les techniques d'imprimerie sont telles que les photos sont généralement tramées. C'est à dire qu'elles sont représentées par une grille régulière de points. Or le caractère récurrent de l'information a des effets pervers... En effet, tout échantillonnage (à la capture ou à l'affichage) dont la fréquence est inférieure à deux fois la fréquence des trames a pour résultat un effet moiré (même phénomène que quand deux rideaux très fins sont l'un sur l'autre). Pour éviter cela, il est donc nécessaire d'acquérir le document avec une résolution suffisante (au moins deux fois celle de la trame) et de "diluer" les point de la trame afin qu'ils occupent toute la surface laissée vierge autour d'eux. Ce traitement appliqué avec un facteur de dilution trop important ou à des documents non tramés donne un résultat flou.

### **c. "Détournage"**

Ce traitement des plus basiques, est en général réalisé à la main, il consiste à ne conserver que la partie des informations contenues dans une zone géométrique (blocs de texte, de figure...), ceci afin de supprimer le superflu (marges, autres blocs...).

### **d. Compression avec perte**

Les méthodes de compression avec perte reposent sur le passage du domaine spatial au domaine fréquentiel : l'image originale est alors considérée comme la somme d'une infinité de composantes dont les premières définissent les grands aplats de couleur, et les suivantes définissent les détails avec une précision croissante. En ne stockant qu'un certain nombre des premières composantes on obtient une bonne approximation de l'aspect général de l'image de départ avec cependant des aberrations à proximité des contours. Lorsque le nombre de composantes augmente, la qualité croît et le rapport de compression décroît. Sur des images sans discontinuités, des taux de compression très élevés peuvent être obtenues sans que les pertes soient visibles à l'œil nu, on recommande donc ce genre de compressions pour des photographies destinées à la visualisation. Par contre, on doit les exclure dans le cas où les documents traités ont des contours très marqués (dessins, textes...), ou qu'ils sont destinés à des traitements automatisés.

### **e. Reconnaissance optique de caractères<sup>3</sup>**

Ce traitement permet d'obtenir à partir de l'aspect optique d'un texte dactylographié, la séquence des caractères présents et jusqu'à un certain point leur mise en page. On sent très bien ici, par la réduction radicale du nombre d'informations pour un même texte que l'on a obtenue des données beaucoup plus riche. Cependant, il faut reconnaître que nous avons perdus tous les caractères inconnus (langues rares, symboles), et que même dans le cas de caractères connus, la reconnaissance est loin d'être infaillible (dans les meilleures conditions : un mot erroné sur vingt).

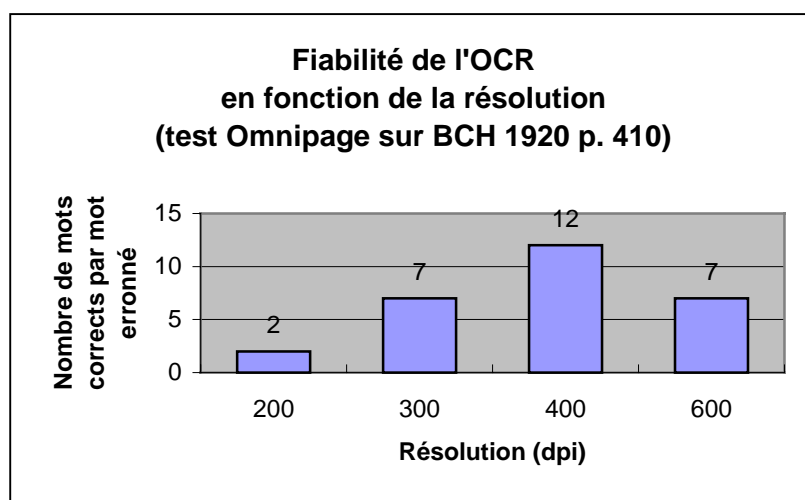
## **3. Un "monde de blocs"**

Après examen de la *Chronique des fouilles*, il apparaît qu'une même page peut rassembler des informations visuellement très différentes : des textes, des gravures, des estampes, des photos. Nous allons, pour chacun de ces types, spécifier les paramètres de numérisation et les traitements à appliquer. Attention, nous nous intéressons ici aux paramètres requis pour une utilisation standard, il y aurait éventuellement possibilité de relier notre projet à un projet de SGBI .

---

<sup>3</sup> En anglais : *Optical Character Recognition [OCR]*

Matériel	Utilisation	Contraintes	Procédé Conseillé
Texte	Traitement par OCR	Optimisation des résultats de l'OCR (contours précis et contrastés, résolution adaptée à la taille des fontes)	Echantillonnage en noir et blanc à 400 dpi ( <i>cf.</i> histogramme ci-dessous) sans "détramage" Compression sans pertes
	Visualisation via le WWW	Texte compris en largeur dans l'écran Lisibilité des lettres Vitesse de transfert	Echantillonnage en noir et blanc à 400 dpi sans "détramage" Conversion en gris Réduction à 200 dpi avec lissage Compression sans pertes
Gravure monochrome	Visualisation via le WWW	Visibilité des détails Vitesse de transfert	Echantillonnage en gris à 200 dpi sans "détramage" Compression sans pertes
Photographie et Gravure en dégradé de gris	Visualisation via le WWW	Rendu des valeurs de gris Vitesse de transfert	Echantillonnage en gris à 200 dpi avec "détramage" Compression à pertes



Le matériel graphique de la chronique, dont les natures sont très diverses, nécessite une numérisation et des traitements adaptés à chacune. Ainsi, il semble impossible, sans entraîner une grosse dégradation de la rentabilité de la numérisation, de traiter indifféremment les différents blocs d'informations d'une seule et même page.

## C. UN NOUVEAU SUPPORT POUR L'ÉCRITURE

Dès les débuts de l'informatique, le souhait est apparu d'utiliser les ordinateurs (que les Anglo-saxons, plus réalistes que nous, appellent "calculateurs") autrement que pour leur fonction originelle, le traitement et le stockage de nombres... L'enjeu était de taille : il s'agissait de l'utiliser comme nouveau support de l'écriture.

Dès lors, il devenait nécessaire de définir des formats électroniques, c'est à dire des règles de codage de ces informations (contenues habituellement dans l'écriture) en une séquence de nombres. La multitude des formats qui se mirent à éclore devint le cauchemar des créateurs et des lecteurs de documents électroniques. Les problèmes posés par cette diversité ainsi que les solutions existantes seront étudiées dans notre première partie.

Dans cette diversité de formats, au-delà des différences de codage, se cachent souvent des différences beaucoup plus profondes. En effet, l'écriture sur papier est héritière d'une longue tradition : des conventions entre créateurs et lecteurs véhiculent implicitement des informations. Que faire de ces conventions ? S'agit-il de retrouver la sémantique sous-jacente et de la coder avec des conventions plus adaptées au nouveau support ? Ou bien s'agit-il de coder simplement les conventions traditionnelles et de laisser le lecteur habitué interpréter ces informations implicites ? C'est cette opposition entre stockage du contenu ou de l'aspect qui sera abordé dans notre seconde partie.

Dans une dernière partie, nous essaierons de voir dans quels cas de figure, l'une ou l'autre des voies considérées précédemment est préférable.

### 1. Formats : le cauchemar de la multitude et de la normalisation

Le constat est là : les formats de documents électroniques sont aussi nombreux que les logiciels qui ont pour mission de les créer, et ceci pour des raisons variées. Tout d'abord, chaque logiciel s'adresse en général à un nouveau besoin et les formats préexistants ne peuvent les satisfaire. Ce phénomène est d'autant plus accru que les éditeurs de logiciels, avides de parts de marchés, inventent constamment pour leurs utilisateurs des nouveaux besoins qui tiennent souvent plus du gadget que de l'indispensable. De plus, il est souvent plus facile de définir ses propres conventions que de tenter de comprendre les conventions des autres, mal documentées ou tout au moins obéissant à un autre mode de pensée. Enfin, la définition de formats propriétaires est souvent une forme de protectionnisme de la part des éditeurs de logiciels dans le but de conserver un monopole.

Devant cette "Tour de Babel" des formats de documents, se pose l'éternel problème de la communication<sup>4</sup>. En effet pour s'assurer à la fois que le document pourra être consulté ou modifié par la plupart de ses destinataires, et qu'il survivra à la disparition du logiciel qui l'a créé, il semble nécessaire d'adopter des standards. Ici, on doit distinguer deux formes de standards...

Un standard *de facto*, souvent un format propriétaire correspondant à un logiciel en situation de quasi-monopole. Non seulement il est utilisable par les possesseurs, en grand nombre, du logiciel en question, mais encore les éditeurs des logiciels concurrents, sous la pression de leurs clients, sont obligés de fournir des filtres permettant de traduire leur propre format dans le format pris comme standard<sup>5</sup>.

Une norme définie par un organisme indépendant (par exemple l'ISO : *International Organisation for Standardisation*) ou par un consortium d'entreprises est souvent le fruit d'une longue recherche afin de répondre à l'ensemble des besoins présents et à venir, et d'un grand souci d'abstraction (laissant de côté les problèmes d'implantation qui par nature différeront d'un système à l'autre). Si ces normes sont très satisfaisantes au point de vue intellectuel, leur mise en pratique est souvent difficile. Tout d'abord, les délais d'élaboration et de publication des normes sont difficilement acceptables dans un marché où les outils se renouvellent continuellement (souvent de manière artificielle dans un but commercial). Ensuite, les questions d'implantation ayant été occultées par la norme, c'est au programmeur du logiciel que revient la tâche laborieuse de faire correspondre ce que la norme prévoit avec ce que la machine peut faire. Enfin, les possibilités offertes par la norme dépassent souvent de

---

<sup>4</sup> Les services informatiques de support aux usagers connaissent bien les questions angoissées du type : « Comment modifier le travail de mon collaborateur qui travaille avec "Claris Works" alors que je n'ai que "Microsoft Word" au bureau ? » ou encore « Puis-je visualiser le rapport de mon prédécesseur créé avec "Word Star" ? ».

<sup>5</sup> Par exemple, le logiciel "Frame Maker" sous UNIX reconnaît (pour la lecture et l'enregistrement) les documents au format "Microsoft Word".

beaucoup les besoins des éditeurs de logiciels qui sont alors tentés pour des impératifs budgétaires de n'implanter qu'une partie de la norme<sup>6</sup>.

Les difficultés que nous venons d'évoquer pour la standardisation poussent beaucoup d'organismes publics à adopter une politique du « Rien ne sert de courir, il faut partir à point », et d'attendre ainsi la définition de normes unanimement reconnues, avant de commencer les vastes projets de transfert de connaissance. Cette attitude prudente ne semble cependant pas être souhaitable<sup>7</sup> si l'on en croit l'exemple de l'histoire récente d'Internet. En effet c'est de la diversité des solutions, des tests "grandeur nature" et du développement par incréments, que sont nés les consensus qui font le succès d'Internet. De plus malgré l'hétérogénéité des solutions, leur utilisation conjointe a pu être maintenue par la création de ponts et de passerelles. Ces notions, valables pour les réseaux, sont largement transposables dans l'ensemble des domaines de l'informatique et en particulier celui des formats de documents électroniques. Inutile donc d'attendre la norme idéale qui satisfera toutes les attentes. Si un format est largement utilisé, cela suffit pour qu'il soit susceptible d'être adopté dans un projet. Il est cependant nécessaire de s'assurer régulièrement qu'il reste consultable par les utilisateurs, et dans le cas contraire de le convertir en un format plus actuel, en utilisant un filtre du marché ou développé pour l'occasion.

En fin de compte, le choix d'un mode de codage plutôt qu'un autre ne doit pas monopoliser l'attention du concepteur d'un système de documents électroniques, étant donné qu'il est toujours possible de créer un automate de conversion. La réelle question n'est pas « Comment coder les informations d'un document électronique ? », mais « Quelles informations coder dans un document électronique ? ».

## 2. Signifiant/Signifié : Que stocker ?

L'écriture repose sur un ensemble de conventions tacites entre le créateur et le lecteur, ceci afin de faire la relation entre le signifiant et le signifié. L'avènement d'un nouveau support pour l'écriture remet en question certaines de ces conventions. Ces remises en question ne sont pas totales : elles ne touchent ni la correspondance mot/idée, ni même la relation alphabet/phonème. Par contre, elles peuvent toucher les correspondances glyphe/alphabet ou encore typographie/structure. Stocker le signifié plutôt que le signifiant représente une plus-value considérable. En effet, une telle solution ne requiert plus l'interprétation du lecteur, et permet par conséquent des traitements automatisés. De plus, dans le cas d'un document créé originellement sous forme électronique, une telle solution libère le créateur des détails de représentation en automatisant complètement le passage du signifié au signifiant, et en garantissant l'homogénéité et la normalisation de cette transition. Par contre dans le cas de la transformation sous forme électronique d'un document existant déjà sous forme papier, cette interprétation a un prix car elle fait appel à l'intelligence d'un opérateur humain ou d'un système expert.

### a. Glyphe/Alphabet

Le lecteur d'un document écrit fait une association mentale entre un dessin et une lettre de l'alphabet. La question est de savoir si l'on enregistre dans un document électronique l'image du glyphe ou bien une référence à la lettre de l'alphabet représentée.

#### (1) Valeur pour l'utilisation

La solution visant à représenter une référence à une lettre plutôt qu'une image du glyphe permet d'automatiser la recherche de mots (exacte ou floue) et d'expressions régulières (définition formelle d'un ensemble de mots satisfaisant des règles données de construction), elle autorise également l'édition du texte (correction, récupération pour citation...), enfin elle rend possible l'adaptation de la représentation aux besoins de l'utilisateur (par exemple à la résolution de son périphérique de visualisation).

#### (2) Coût de la création *ex nihilo*

La solution évoquée ci-dessus se traduit par une réduction du coût lorsque le document est créé dès l'origine sous forme électronique, car le dactylographe n'a pas à se soucier de l'aspect des caractères (police, taille) mais uniquement des lettres. Le choix ou la modification de l'aspect pourra être réalisée plus tard par le metteur en page ou même par le lecteur, et ceci en une seule opération pour l'ensemble du document.

---

<sup>6</sup> Par exemple, pour le format TIFF, format de représentation d'une image numérique, il n'existe, à ma connaissance, aucun logiciel du marché capable de reconnaître l'intégralité de ses spécifications.

<sup>7</sup> CHARITY, M. N. Multiple Standards? No problem. *Digital Libraries '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station (Texas), June 19-21 1994* [On-line]. Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/DL94/position/charity.html>>

### (3) Coût de l'interprétation à partir d'un document existant

Dans le cas de la transformation d'un document existant (sous forme papier), l'interprétation nécessaire pour une telle solution peut soit être réalisée par un être humain, lisant les pages, identifiant les lettres, et tapant les touches correspondantes d'un clavier, soit être réalisée par un automate de reconnaissance optique de caractère (en anglais : Optical Character Recognition [OCR]). Cette interprétation s'accompagne invariablement d'erreurs (d'avantage pour l'automate que pour le dactylographe), pouvant être minimisées par l'utilisation de correcteurs orthographiques (choix manuels parmi les propositions d'un automate). Le coûteux processus de la relecture par un expert reste néanmoins le seul garant de la qualité d'une publication.

### (4) Problèmes inhérents et solutions

Faire référence à des lettres de l'alphabet suppose qu'un choix préliminaire ait été réalisé sur les lettres à utiliser et les codes les représentant. Un même nombre peut ainsi correspondre à différentes lettres suivant la norme utilisée, ceci représentant une réelle entrave aux échanges de documents notamment entre pays ayant des besoins différents (signes diacritiques, alphabets, idéogrammes). Pour éviter cette translittération involontaire (transcription de chaque signe d'un système d'écriture en un signe d'un autre système), l'adoption d'une norme internationale<sup>8</sup> répondant aux besoins de tous semble être la solution la plus satisfaisante mais c'est aussi la plus difficile à intégrer dans des logiciels ; d'autres solutions à court terme ont été utilisées comme par exemple l'adoption unanime d'une norme minimale (signes anglo-saxons<sup>9</sup>), et la transcription des caractères spéciaux en une séquence donnée de caractères appartenant à cette norme minimale<sup>10</sup>. Un autre problème réside dans le besoin qu'à un document de contenir des signes rares ou pour lesquels on souhaite retranscrire l'aspect et non la signification (figures, caractères de langues oubliées, copie fidèle d'un original). On mémorise alors l'image du signe sous forme d'une matrice de points de différentes couleurs. Deux solutions s'offrent alors : soit les nombres représentant cette matrice sont transcodés en caractères et insérés dans le texte, soit ces nombres sont stockés sous leur forme native dans un fichier à part et une référence à ce fichier est placée dans le texte à l'endroit où devrait se placer le signe.

## **b. Typographie/Structure**

Un document écrit n'est pas uniquement une séquence de mots, il est en général fortement structuré. L'existence de titres, de notes, de paragraphes, de vers, de blocs de citation, de dialogue, sont autant de témoins de ces informations structurales. Elles sont nécessaires tant pour retrouver un passage que pour connaître le contexte qui donne son véritable sens aux mots.

Ces informations sont traditionnellement indiquées sur papier par des conventions typographiques : marges, attribut des caractères (gras, italique, souligné, taille), alignement (centré, à droite, justifié), etc. Doit-on lors de la création de documents numériques mémoriser ces indications typographiques ou bien leur signification ? Le problème est comparable au précédent s'agit-il de laisser au lecteur le soin d'interpréter un certain symbolisme ou bien doit-on effectuer l'interprétation dès la création du document électronique afin d'utiliser toute la puissance de traitement des automates ?

### (1) Valeur pour l'utilisation

La mémorisation de la structure permet, par exemple, au lecteur de "naviguer" parmi l'organisation hiérarchique d'un document et de trouver directement un passage plutôt que de parcourir l'ensemble du texte. Elle permet également d'affiner considérablement la recherche par listes d'occurrences : un même mot n'a en effet pas la même importance pour un document s'il est présent dans son titre principal, un de ses sous-titres, dans une note de bas de page ou dans le corps d'un paragraphe.

### (2) Coût de la création *ex nihilo*

Ici encore, la mémorisation du signifié plutôt que du signifiant se traduit, lors de la production *ex nihilo* d'un document électronique, par une baisse des coûts. Inutile pour le dactylographe d'avoir une connaissance exhaustive des règles typographiques nationales (définies par exemple par l'Imprimerie Nationale) ou d'en définir de nouvelles au fur et à mesure de l'élaboration du document (avec la difficulté de correction et les risques d'hétérogénéité que cela peut entraîner). Peu importe de savoir si tel titre doit être dans une police sans-serif ou romaine, en gras ou en italique, précédée d'un chiffre romain ou arabe, si ce chiffre doit être un quatre ou un cinq ; il suffit de savoir qu'il s'agit d'un titre de niveau 2. Si la présentation sous la forme typographique est nécessaire, un automate se chargera aisément de la conversion conformément à des règles données.

---

<sup>8</sup> UNICODE, <http://www.unicode.org>

<sup>9</sup> American Standard Code for Information Interchange [ASCII].

<sup>10</sup> Par exemple, le caractère diacritique français " é " est codé " \ ' e " en langage T<sub>E</sub>X et " &eacute; ; " en HTML.

### (3) Coût de l'interprétation à partir d'un document existant

La traduction de la typographie d'un document papier en indications structurales n'est pas un processus trivial. En effet deux blocs de même typographie ne sont pas forcément de même type, structurellement parlant. Par exemple, le résumé présent au début des articles scientifiques présente la même typographie qu'un bloc de citation<sup>11</sup>. Dans ce cas précis, c'est le fait que l'un soit situé avant l'introduction qui permettra de l'identifier comme étant un résumé. Ainsi, il existe toujours un ensemble de règles permettant de passer du signifiant au signifié (celles-là mêmes que nous appliquons sans être capable de les exprimer), mais elles ne sont pas évidentes. Ici encore, cette interprétation nécessite le concours d'un être humain ou d'un système expert.

### (4) Problèmes inhérents et solutions

Une telle politique n'est applicable que pour des documents dont la typographie a un sens profond et non une seule raison esthétique (affiches publicitaires, etc.). On serait tenter de penser que, même pour un livre, certains attributs de mise en page n'existent que pour eux-mêmes sans avoir de sens particulier et qu'il serait nécessaire d'intégrer, dans notre codage, des indications sur l'aspect du document. Il apparaît en fait que ces attributs typographiques n'ont soit pas de raison d'être, soit corresponde à des structures très fines. Ainsi, par exemple, tel mot n'est pas en italique pour être en italique, mais parce qu'il s'agit d'une expression latine. Se pose ici le problème du raffinement du modèle structurel<sup>12</sup>. La plus-value du document sera d'autant plus élevée que le modèle sera fin. Par exemple, il est utile de savoir que tel bloc est une note de bas de page, mais il est plus intéressant de savoir que telle partie du bloc est une référence bibliographique, et il est inestimable de savoir que telle fraction de la référence est le titre d'un ouvrage... Un choix difficile se pose alors : doit-on adopter un modèle général préexistant ou bien développer un modèle propre au corpus de documents ? L'adoption de modèles largement reconnus<sup>13</sup> présente l'avantage de disposer de logiciels existants (notamment pour l'affichage sous une forme typographique) et de permettre des traitements automatisés (recherches...) sur plusieurs corpus à la fois. Cependant, un modèle général sera souvent insuffisamment fin pour représenter la richesse structurelle des documents du corpus. Il est donc indispensable de vérifier, avant d'utiliser un modèle général, son adéquation avec le corpus de documents considéré.

Une autre entrave réside dans le fait que, par la force de l'habitude, nous ne distinguons pas la forme imprimée d'un document d'avec son contenu. Ainsi, un paragraphe sera souvent référencé par un numéro de page pourtant susceptible de changer d'une édition à une autre, ne contenant aucune indication sur le sujet traité, et ne correspondant que très partiellement au paragraphe (le paragraphe pouvant s'étendre sur plusieurs pages et la page pouvant contenir plusieurs paragraphes), alors qu'il serait grandement préférable de le référencer par son "chemin hiérarchique". Par exemple, une référence au présent passage au lieu d'être sous la forme :

Etude globale du projet, p. 15

devrait s'écrire :

```
Etude globale du projet
  Rapport scientifique
    Un nouveau support pour l'écriture
      Signifiant/Signifié : Que Stocker ?
        Typographie/Structure
          Problèmes inhérents et solutions
```

Nous ne pouvons donc pleinement profiter de la représentation sémantique d'un document que lorsque les habitudes auront changé et que notamment les normes bibliographiques auront évolué dans ce sens. En attendant, il sera nécessaire de conserver à l'intérieur du document électronique des informations (telles que le numéro de page) relatives à son aspect sous une de ses formes papier, perdant ainsi une partie du bénéfice du pouvoir d'abstraction de la représentation.

<sup>11</sup> **KASDORF, Bill.** SGML and PDF, Why We Need Both. *Journal of Electronic Publishing*, June 1998, Volume 3, Issue 4 [On-line].

Available on Internet <URL : <http://www.press.umich.edu/jep/03-04/kasdorf.html> >

<sup>12</sup> Par exemple, ce genre de modèles structurels est appelé *Document Type Definition* [DTD] dans le jargon du *Standard Generalized Markup Language* [SGML].

<sup>13</sup> Par exemple, un projet de recherche international (appelé *Text Encoding Initiative* [TEI]) a élaboré un DTD pour les publications en humanité.

### 3. Quels choix pour quels besoins ?

Dans le but de clarifier la situation, au risque de paraître catégorique, j'aurais envie de dire que les solutions basées sur l'aspect du document seraient plutôt tournées vers le passé, et celles basées sur le contenu sémantique plutôt vers le futur.

#### **a. L'héritage du passé**

L'expression "vers le passé" n'a ici rien de péjoratif, il s'agit uniquement de rappeler que nous sommes héritiers d'une longue tradition du livre, et que par conséquent nos modes de pensées et nos techniques sont fortement dépendantes de l'objet physique "livre".

En effet, d'une part malgré les titres illusoires des journaux, nous ne vivons pas dans une société du multimédia mais dans une société du papier. Même, lorsqu'un document est créé par ordinateur, il finit toujours, tôt ou tard, à être matérialisé par une imprimante ! Nous sommes plus accoutumés au papier que nous le croyons (même les informaticiens !) : notre vitesse de lecture est plus élevée sur un papier que sur un écran, on se perd encore dans le labyrinthe des liens hypertextes, on ne profite que de manière insuffisante de la puissance du "copier-coller" et des liens dynamiques entre documents, et enfin notre façon de penser à un document est complètement empreinte de l'influence de sa forme papier (référence à des numéros de page...).

D'autre part, notre époque se trouve face à une tâche titanesque : celle de numériser (au sens large) l'ensemble de l'héritage écrit qui nous soit parvenu. Or nous avons vu que si la numérisation de l'aspect des documents peut être facilement automatisée, celle du contenu sémantique l'est beaucoup plus difficilement. De fait, la plupart des solutions commerciales pour ce transfert de connaissance s'en sont tenues à l'aspect des documents, venant encore accentuer la différence de difficulté des deux voies.

#### **b. Vers les documents du futur**

Il est évident que le futur des documents électroniques sera la conservation d'informations sémantiques et non d'aspect. En effet, seule la conservation du sens permettra de profiter pleinement de la capacité de traitement des automates (recherche d'informations, résumé, etc.). De plus, une fois la phase de traduction papier/numérique passée, lorsque les documents seront créés dès le départ sous forme électronique sémantique, cela se manifestera par une baisse du prix et une plus grande homogénéité de la production des publications.

#### **c. Et le présent dans tout ça ?**

« Oui mais que faire *maintenant* pour numériser nos documents papiers ? » me direz-vous. Effectivement, nous nous trouvons réellement à une époque charnière. Doit-on profiter des acquis du passé ou anticiper sur l'avenir ? Difficile question, je l'avoue. D'un côté, pour la numérisation de l'aspect, nous possédons toute la technologie nécessaire (matérielle et logicielle) pour un prix raisonnable, nous savons que les utilisateurs possèdent toutes les connaissances et les automatismes pour interpréter cet aspect. De l'autre côté, pour la numérisation du contenu, de coûteuses interventions humaines sont nécessaires, les systèmes experts qui pourraient s'en charger restent à faire où sont en cours de réalisation<sup>14</sup>, les utilisateurs sont très peu préparés à la consultation de ce genre de documents, et nous savons pourtant malgré tout cela que le progrès passera par-là....

Je pense que le choix doit se faire par rapport à la nature profonde du projet. S'il s'agit d'un projet d'application, à moyen terme, aux ressources financières limitées, dans laquelle l'obligation de réussite se situe au niveau de la satisfaction des besoins des usagers actuels, la numérisation de l'aspect sera la solution la plus raisonnable. Par contre, s'il s'agit d'un projet de recherche, à long terme, aux ressources financières importantes, dans lequel l'obligation de réussite se situe surtout dans l'exploration d'une solution, il est certain que la numérisation du contenu représente de loin la voie la plus intéressante.

En résumé, nous avons mis en évidence le fait que malgré la multitude des formats de documents électroniques et les problèmes qu'elle engendre, ces problèmes ne sont pas insurmontables tant que la différence des formats se situe au niveau du codage. Par contre, il existe une barrière beaucoup plus fortes entre deux classes de documents électroniques : ceux qui mémorisent l'aspect du document et ceux qui mémorisent son

---

<sup>14</sup> PALOWITCH, Casey and STEWART, Darin. Automating the Structural Markup Process in the Conversion of Print Documents to Electronic Texts. *Second Annual Conference on the Theory and Practice of Digital Libraries, Austin (Texas), June 11-13 1995* [On-line]. Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/DL95/papers/palowitc/palowitc.html> >

contenu. Il s'agit donc de faire un choix entre les deux, choix qui risque d'être en partie irrévocable. Il apparaît que c'est la mémorisation du contenu (que ce soit au niveau des lettres de l'alphabet ou de la structure du document), même si elle anticipe l'adaptation des utilisateurs et les moyens technologiques, qui apportera une plus-value supérieure au document dans le futur puisqu'elle permettra l'utilisation d'automates pour de nombreux traitements. Un projet doit bien sûr se réclamer actuellement du domaine de la recherche à long terme pour prendre une telle option.

## D. RECHERCHE D'INFORMATIONS TEXTUELLES

Que serait une bibliothèque dont les livres ne seraient pas classés, dont il n'existerait aucun inventaire, et dont personne ne connaîtrait le contenu ? Un vain bibelot... Il en est de même pour une bibliothèque virtuelle. Aussi grandes seraient la quantité et la qualité de ses textes, la collection serait inutilisable si l'on ne disposait pas de techniques efficaces de recherche d'informations. En effet, le chercheur, spécialiste de son domaine, s'intéresse seulement à certaines parties d'ouvrages. Il attend, face à une question précise, de voir comment elle a été traitée et quelles ont été les réponses apportées antérieurement. La recherche d'informations consiste donc à associer à chaque question des chercheurs d'informations un ensemble d'extraits (plus ou moins fins) de textes du corpus répondant au mieux à la question. Afin de définir les moyens humains et techniques à prévoir dans une bibliothèque virtuelle, nous nous intéresserons ici à deux types de recherche d'informations. Dans une première partie, nous aborderons les méthodes exactes, basées sur la compréhension intellectuelle des textes et des questions, et dans une deuxième partie nous traiterons des méthodes approchées, basées sur des critères statistiques.

### 1. Recherche intellectuelle

La recherche d'informations, telle que nous l'avons définie précédemment comme mise en correspondance d'une question et d'un extrait d'ouvrage, laisse supposer la nécessité d'un agent, humain ou automatique, qui aurait connaissance du contenu des ouvrages et de la question, qui en aurait compris le sens, et qui les aurait mis en rapport. Nous allons dans cette partie mettre en évidence les formes, les forces et les faiblesses que peut prendre ce service selon qu'il est effectué par un humain ou un automate.

#### a. « De bouche à oreille »

La méthode optimale pour le chercheur d'informations, consiste à demander à des spécialistes (bibliothécaires ou autres chercheurs). En effet, eux seuls sont capables de comprendre le contenu sémantique des ouvrages, de porter dessus un jugement critique, d'en faire une synthèse, de faire des recoupements avec d'autres ouvrages, de le compléter par des informations informelles (non - officielles, subjectives), de comprendre la question du chercheur d'informations, de l'aider si besoin à raffiner ou à étendre sa requête. Il est donc indispensable, dans le cadre de la conception d'une bibliothèque virtuelle, de ne pas négliger les moyens d'assurer ces interactions sociales<sup>15 16</sup> (chercheur/chercheur, chercheur/bibliothécaire) sous peine de voir disparaître un des fonctions primordiales des bibliothèques.

#### b. Anticipation

Si une telle expertise humaine est incomparable, il s'agit malheureusement d'une "ressource critique", le nombre d'experts étant faible, leur disponibilité réduite, leurs services coûteux. Il est en effet impensable d'imaginer un bibliothécaire derrière chaque usager ! C'est pourquoi, il existe dans les bibliothèques traditionnelles des index thématiques (avec vocabulaire contrôlé), des résumés, des dossiers de presse, des livres mis en évidence (nouveau), des notes bibliographiques, des études critiques... Tous ces procédés reposent sur la même idée : celle d'anticiper les besoins des usagers de la bibliothèque afin d'utiliser une seule fois l'expertise des spécialistes (bibliothécaires et chercheurs) et de la partager entre les utilisateurs aussi souvent que nécessaire. Il est à noter que les bibliothèques virtuelles ont tout à gagner à transposer et adopter ces méthodes traditionnelles, en particulier l'indexation. Toutefois l'indexation est une tâche particulièrement complexe, coûteuse, et subjective<sup>17</sup>. De plus, on peut se demander s'il est réellement rentable de réaliser ces index dans le

---

<sup>15</sup> ACKERMAN, Mark S. Providing Social Interaction in the Digital Library. *Digital Libraries '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station (Texas), June 19-21 1994*. [On-line].

Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/DL94/position/ackerman.html>>

<sup>16</sup> TOCHTERMANN, Klaus. *A First Step Toward Communication in Virtual Libraries*. [On-line]. College Station (Texas), Texas A&M University, Center for the Study of Digital Libraries, 1996.

Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/pubs/klaus/TecRepKlaus.html>>

<sup>17</sup> DAVID, C., GIROUX, L., BERTRAND-GASTALDY, S., et al. Indexing as Problem Solving: a Cognitive Approach to Consistency. *ACSI 95: Annual Conference of the Canadian Association for Information Science, Edmonton (Alberta), June 7-10 1995*. [On-line].

Available on internet : <URL : <http://www.ualberta.ca/dept/slis/cais/david.htm>>

cas de la recherche scientifique (science humaine ou de la nature) qui par essence prospecte à chaque fois des horizons inexplorés et différents. Une méthode économique consisterait à créer les clefs au fur et à mesure de la demande, on peut même prévoir d'impliquer l'utilisateur dans la création de l'index (par exemple en récupérant son environnement paramétrable). A ce sujet, nous devons cependant émettre une réserve, car certaines études affirment que les chercheurs en sciences humaines seraient beaucoup plus individualistes que les chercheurs en sciences de la nature<sup>18</sup>. Reste à définir dans quelles mesures cet individualisme pourrait être gênant et comment il pourrait être contourné.

### **c. Automatisation ?**

Il est tentant, devant une tâche aussi coûteuse, de s'interroger sur l'aide que pourrait fournir un automate... Cependant, un automate ne peut traiter que de problèmes complètement formalisés et il n'en est rien : trouver une référence bibliographique en réponse à une question nécessite la formalisation du langage naturel des ouvrages et de la question. Or, après l'euphorie des années 1970<sup>19</sup>, où l'on pensait découvrir la "structure profonde" du langage et par-là le moyen de passer automatiquement du langage naturel à un langage formel, les travaux de recherche ont mis en évidence l'extrême complexité du langage naturel<sup>20</sup> (sous-entendu, quasi-infinité de constructions), qui rend même mystérieuse son apprentissage par les êtres humains.

Une autre solution consisterait à demander aux utilisateurs de poser leurs questions dans un langage formel, à des experts de formaliser le contenu des ouvrages, et à l'automate de trouver les correspondances entre les deux. Outre tous les problèmes de formation humaine que cela comporte, la traduction en langage formel de plusieurs milliers de page est impensable. En effet, même en supposant que tout ce qui est exprimé en langage naturel puisse l'être dans un langage formel (ce qui est loin d'être évident), une telle traduction est terriblement longue et difficile : c'est ce que font tous les jours les concepteurs de logiciels lorsqu'ils créent un modèle à partir d'un document de spécification, mais ces derniers ne font que quelques pages...

La recherche d'informations basée sur le sens du texte reste donc le domaine réservé de l'expert humain, l'automate s'avérant ici assez inapte. Toutefois pour ne pas abuser du temps précieux de ces experts, il est nécessaire de mettre en place d'autres techniques complémentaires de recherche, même moins efficaces, mais qui puissent être automatisées.

## **2. Recherche statistique**

Devant l'impuissance des automates à traiter les textes comme des collections de connaissances, on en est réduit à utiliser les automates pour traiter les textes comme de simples séquences de caractères. C'est pourquoi l'une des applications informatiques les plus courantes pour les textes intégraux consiste à rechercher les occurrences d'une chaîne de caractères. L'automate s'acquitte fort bien de cette tâche, mais l'utilisateur s'intéresse en général plus à une idée qu'à un mot. C'est ainsi que se posent deux problèmes à savoir le "bruit" et le "silence". Le silence correspond aux cas où des documents qui auraient pu répondre à la question de l'utilisateur n'ont pas été choisis par l'automate. Le bruit, quant à lui, correspond aux documents choisis par l'automate alors qu'ils ne correspondent pas à la question.

### **a. Crever le silence**

La notion d'égalité de deux séquences de caractères est terriblement restrictive. Aussi ont été mises en place de nombreuses techniques afin de l'étendre<sup>21</sup>.

La première cause de silence est sans doute le polymorphisme des mots. Il est nécessaire, par exemple, de faire des comparaisons indépendantes de la casse (majuscule/minuscule). On peut également rechercher un

---

<sup>18</sup> **BEGUIN, Daniel.** Les antiquisants face à l'informatique et aux réseaux. *Internet et les chercheurs - Rapport intermédiaire*. [En ligne]. Paris, Ecole Normale Supérieure, Département de Sciences Sociales, Novembre 1996. Disponible sur internet : <URL : <http://elias.ens.fr/atelier/articles/ArticleInternetnov96.html>>

<sup>19</sup> **LESK Michael.** *Seven Ages of Information Retrieval*. [On-line]. Ottawa, International Federation of Library Associations and Institutions, Universal Dataflow and Telecommunications Core Programme, March 1996.

Available on internet : <URL : <http://www.ifla.org/ifla/VI/5/op/udtop5/udt-op5.pdf>>

<sup>20</sup> Cf. les travaux de **CHOMSKY, Noam**.

<sup>21</sup> **MCKINLEY Tony.** *From Paper to Web*. [On-line]. Indianapolis (Indiana), Adobe Press, 1997. Chapter 12, Advanced Searching Techniques.

Available from internet : <URL : [http://imagebiz.com/paperweb/ptweb\\_12.pdf](http://imagebiz.com/paperweb/ptweb_12.pdf)>

ensemble de formes d'un même mot en utilisant des expressions régulières<sup>22</sup> ou des caractères jokers<sup>23</sup>. La maîtrise de tels formalismes étant un peu difficile, une solution consiste à demander à l'automate d'obtenir la racine du mot en supprimant les préfixes et suffixes et de sélectionner tous les mots ayant la même. Afin d'éviter les silences dus à des erreurs de frappe ou de reconnaissance optique de caractères [OCR], on peut également utiliser des "N-grams"<sup>24</sup> (ou logique floue), c'est à dire mettre en relation des suites de caractères qui, bien que différentes, se ressemblent de par les caractères qui les composent et leur enchaînement.

La deuxième cause de silence vient du fait que pour une même idée, il existe plusieurs mots, ceci dans le cadre d'une synonymie stricte ou dans des relations plus larges comme la généralisation ou l'analogie. La meilleure solution repose sur l'utilisation de larges ressources lexicales spécifiques au domaine d'étude du corpus<sup>25</sup>. Une solution moins coûteuse, mais plus approximative, consiste à utiliser le corpus lui-même comme représentatif des relations sémantiques entre les mots d'après leurs relations de proximité<sup>26</sup>.

## **b. Etouffer le bruit**

Si les méthodes que nous venons d'évoquer nous permettent d'atteindre une forte proportion des textes en rapport avec la question, elles présentent cependant le sérieux inconvénient de générer un certain nombre de réponses qui n'ont que peu de rapport avec la question<sup>27</sup>. Il devient donc nécessaire de séparer au mieux, parmi ces réponses, le grain de l'ivraie...

La première catégorie de méthodes consiste à spécialiser la question afin de supprimer au mieux les réponses non pertinentes. On peut ainsi restreindre la recherche aux textes comportant plusieurs expressions à la fois, avec une proximité donnée ou dans un même paragraphe, ou bien encore en omettant les textes qui comportent des termes non désirés. Ces méthodes peuvent s'appliquer à des textes bruts. Par contre, d'autres méthodes plus puissantes s'appuient sur des informations supplémentaires relatives au texte (appelées pompeusement "méta-données"). On pourrait ainsi restreindre la portée d'un terme qui existerait dans plusieurs domaines d'étude<sup>28</sup> à condition de disposer du sujet du texte. De même, dans un texte balisé<sup>29</sup>, on pourrait limiter la recherche d'occurrence d'un terme à un contexte particulier<sup>30</sup>.

La deuxième catégorie de méthodes considère la pertinence comme une notion relative. De fait, il n'y a pas, comme précédemment, suppression de certaines réponses, mais classement des réponses par ordre de pertinence. L'évaluation de cette pertinence, ne pouvant se faire sur des critères sémantiques, se fait sur des critères dont l'adéquation reste approximative : nombre de correspondances, densité de ces correspondances, date d'édition, importance relative des éléments de la question... Par rapport à ce dernier point, il est à noter qu'il est souvent nécessaire, en effet, d'affecter plus ou moins d'importance aux différents termes d'une question. Ceci peut soit se faire de manière explicite par l'utilisateur, soit être réalisé par l'automate en fonction du corpus. En effet, on suppose

---

<sup>22</sup> Par exemple : " scénari[oi]" avec une syntaxe donnée peut désigner l'ensemble des mots "scénario" et "scénarii".

<sup>23</sup> Par exemple : "\*histoire" désigne l'ensemble des mots obtenus en remplaçant l'astérisque par une suite (vide ou non) de caractères : "histoire", "préhistoire", "protohistoire", etc.

<sup>24</sup> **CAVNAR, William B., GILLIES, Andrew M.** Data Retrieval and the Realities of Document Conversion. *Digital Libraries '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station (Texas), June 19-21 1994.* [On-line].

Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/DL94/position/cavnar.html>>

<sup>25</sup> **BANERJEE, Sujata, MITTAL, Vibhu O.**, On the Use of Linguistic Ontologies for Accessing and Indexing Distributed Digital 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.* [On-line].

Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/DL94/paper/banerjee.html>>

<sup>26</sup> **FUTRELLE, Robert P., ZHANG, Xiaolan.** Large - Scale Persistent Object Systems for Corpus Linguistics and Information Retrieval. *Digital Libraries '94: Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station (Texas), June 19-21 1994.* [On-line].

Available on internet : <URL : <http://www.csdl.tamu.edu/csdl/DL94/paper/futrelle.html>>

<sup>27</sup> Par exemple, la recherche par N-gram appliquée à "comte" risque de nous retourner les occurrences de "compte" et "conte" ; de même, la recherche par racine lexicale appliquée à "histoire" pourrait nous retourner les occurrences de "histogramme" !

<sup>28</sup> Par exemple, " le matériel" peut être utilisé autant en informatique (comme l'opposé du logiciel), qu'en philosophie (comme l'opposé du spirituel), qu'en archéologie (comme synonyme de vestige).

<sup>29</sup> Cf. section II. C.

<sup>30</sup> Par exemple, ne chercher les occurrences du nom d'un auteur que dans les références bibliographiques.

qu'un mot est d'autant plus important dans une question, qu'il est précis, et donc qu'il apparaît rarement dans un corpus<sup>31</sup>.

En résumé, nous avons vu qu'une recherche optimale d'informations supposait la compréhension du sens de la question, ainsi que du corpus de textes. Devant l'inaptitude des automates à effectuer de telles tâches, il paraît donc indispensable de prévoir, dans l'élaboration des bibliothèques virtuelles, les moyens d'assurer les interactions entre les bibliothécaires et les chercheurs (index thématiques, résumés, dossiers, réponses aux usagers...) et entre les chercheurs eux-mêmes (notes bibliographiques, études critiques, discussions...). Toutefois l'expertise humaine reste coûteuse, et on est donc amené à compléter ces méthodes exactes de recherche par des méthodes approchées basées non plus sur des critères sémantiques mais statistiques (pouvant être automatisées). Enfin, il est souhaitable d'utiliser conjointement le plus grand nombre de ces techniques de recherche approchée afin d'obtenir la plus grande pertinence possible.

---

<sup>31</sup> Par exemple, si la question était "céramique MR IIIC", les trois mots, de part leur fréquence respective dans un corpus de textes archéologiques, seraient classés du plus spécifique au plus général de la manière suivante : "IIIC", "MR", puis "céramique".

## E. VERS UNE SOLUTION

### 1. Pourquoi des documents électroniques ?

Un document sous une forme numérique est une suite finie de nombres appartenant à un ensemble fini de valeurs. Une telle information présente la particularité de pouvoir être copiée sans perte une infinité de fois pour sa conservation ou sa diffusion, et de pouvoir être traitée directement par ordinateur pour sa consultation ou sa modification. Cela se traduit par une baisse des coûts de distribution, de mise à jour, et par un accroissement de la qualité de service (recherche d'informations...).

### 2. Et les périodiques scientifiques ?

Par rapport à la distribution d'une monographie, celle d'un périodique est un paramètre particulièrement critique, étant donné qu'elle doit avoir lieu principalement avant l'émission du numéro suivant. D'autre part, le discours scientifique, contrairement au discours littéraire, est souvent contraint à faire appel à d'autres langages que notre langue naturelle : s'il s'appuie sur le réel, il fera abondamment usage de photographies, de fac-similés ; s'il tente de le modéliser ou de l'abstraire, il fera appel à des tableaux, des schémas, des équations... Une livraison de périodique scientifique, de plus, s'inscrit en général dans un projet d'ensemble : elle ne doit pas être considérée comme un ouvrage unique mais comme une partie d'un corpus formé par la globalité des livraisons. Enfin, il s'adresse à des spécialistes, intéressés non pas par l'ensemble de la livraison, mais par un ou des passages précis. Aussi, tout est prévu pour une lecture non-linéaire et rapide : organisation hiérarchique, table des matières, des figures, index, bibliographies...

### 3. Que stocker ?

Les termes même de « document » ou de « livre » portent toute l'ambiguïté de la question : s'agit-il d'œuvres intellectuelles, ou bien d'objets que l'on peut parcourir des yeux ? De la même façon, doit-on d'un document conserver le signifiant ou le signifié ? Et ce jusqu'à quel niveau ? Conserver les mots ou le sens ? La typographie ou la structure logique ? Les glyphes ou les caractères alphabétiques auxquels ils font référence ? Il apparaît que plus une information est abstraite moins elle est redondante (d'où un volume plus réduit – et donc une distribution et une conservation moins coûteuse – mais aussi, un gain de productivité dans la création et la mise à jour, une intégrité et une homogénéité plus grande, et enfin la possibilité de mieux simuler par des traitements automatisés des comportements humains). Cette abstraction, que nous appellerons « modélisation sémantique » a cependant un coût, qui ne peut être justifié que par l'utilisation qui est faite du document. En effet, pour un roman, lu en général du début à la fin, un fac-similé (de la version papier) numérique serait tout à fait suffisant. Par contre, comme nous l'avons vu plus haut, un périodique scientifique mérite dans sa version numérique un effort de formalisation comparable à celle dont il a fait l'objet pour sa version sur papier : structure logique, sémantique ou étendue d'application des termes principaux...

### 4. Sous quelle forme ?

Qu'il s'agisse de tables des matières, de thesaurus ou d'autres formes de modélisation d'un texte, on sent *a priori*, qu'il s'agit de structures complexes (graphes, hypergraphes, arbres...) où les mots sont reliés les uns aux autres suivant plusieurs dimensions. Le problème consiste à projeter sans perte ces multiples dimensions en une dimension unique (les anglo-saxons parlent de *serialisation*), pour la stocker dans un fichier numérique. Pour ce faire, nous avons la chance de disposer, du « Langage de Marquage Extensible<sup>32</sup> », norme internationale, appelée à être LE standard pour les informations semi-structurées, et qui fait actuellement l'objet de nombreux développements d'outils logiciels adaptés.

### 5. Du papier au numérique, quelle voie ?

Avant toute chose, nous devons comprendre que les informations sémantiques que nous choisirons de stocker seront de deux ordres : les premières seront déjà en évidence dans l'aspect du document « papier » (par la mise en page, la typographie, etc.) les autres seront présentes seulement de manière implicite dans le texte. Pour les premières, deux voies pourront être envisagées : une voie automatique (si les logiciels de « reconnaissance

---

<sup>32</sup> *Extensible Mark-up Language* [XML], établi par le *World Wide Web Consortium* [W3C].

optique de caractères » du marché extraient assez mal la structure logique d'un document, certains laboratoires de recherche se sont spécialisés dans ce qu'ils appellent la « Dématérialisation de documents », et une voie manuelle (par saisie). Au sujet des coûts respectifs de ces solutions, il est nécessaire de faire une étude dénuée de tous préjugés (la « reconnaissance optique de caractères » peut nécessiter par exemple une coûteuse correction manuelle, la saisie à la main, quant à elle, peut être effectuée à l'étranger pour des coûts raisonnables). Enfin, pour le deuxième type d'information, seule est possible une saisie manuelle effectuée par du personnel comprenant la langue et possédant une certaine culture générale.

## 6. Peut-on parler de bibliothèque virtuelle ?

Une fois que nous posséderons ce corpus de l'ensemble des livraisons d'un périodique, stocké numériquement dans sa forme sémantique, aura-t-on pour autant une bibliothèque virtuelle ? Tout d'abord, la forme sémantique choisie, si elle est bien adaptée à l'ordinateur ne l'est pas pour le lecteur humain, il faudra donc prévoir des programmes de traduction dans des formes exploitables par le lecteur (cf. Figure 1). D'autre part, il est intéressant de remarquer que si la bibliothèque reste la métaphore principale d'un tel système d'information, elle ne suffit pas à la décrire : en effet, nous avons la possibilité, à partir des données contenues dans le corpus, non seulement de recréer les volumes qui existaient sous forme « papier », mais encore de créer, à la demande, l'ouvrage dont a besoin l'utilisateur (par exemple une rétrospective de tous les articles traitant du même sujet). Enfin, si nous souhaitons profiter pleinement de la métaphore de la bibliothèque pour la capitalisation et le transfert de connaissances, il ne faut pas oublier qu'une bibliothèque n'est pas seulement une collection de livres, mais aussi un lieu de rencontres entre personnes. Aussi, la bibliothèque virtuelle ne peut remplir son rôle que si l'on donne les moyens aux gens qui la « fréquentent » d'échanger leurs idées et leurs questions.


<p> <input type="checkbox"/> Bulletin de Correspondance Hellénique, 1920  <input type="checkbox"/> Chronique des fouilles et découvertes archéologiques dans l'orient hellénique (Novembre 1919 - Novembre 1920)  <input type="checkbox"/> Personnel, généralités  <input type="checkbox"/> Fouilles, découvertes, travaux archéologiques  <input type="checkbox"/> Attique  <input type="checkbox"/> Péloponnèse  <input type="checkbox"/> Grèce Centrale  <input type="checkbox"/> Iles Ioniennes, Acarnanie, Epire  <input type="checkbox"/> Eubée  <input type="checkbox"/> Thessalie  <input type="checkbox"/> Iles de l'Egée  <input type="checkbox"/> Macédoine  <input type="checkbox"/> Archipel thrace  <input type="checkbox"/> Thrace  <input type="checkbox"/> <u>Thrace occidentale</u>  <input type="checkbox"/> Constantinople  <input type="checkbox"/> Eléonte de Thrace  <input type="checkbox"/> Thrace orientale  <input type="checkbox"/> Asie-Mineure, îles de la côte d'Asie         </p>	<p style="text-align: right;"><i>BCH</i>, XLIV, p. 409-410</p> <p>La Thrace occidentale ayant été administrée par un corps d'occupation français, d'octobre 1919 au 28 mai 1920 (1), il a été procédé, par les soins des autorités militaires et sur questionnaire fourni par l'Ecole française d'Athènes, à une enquête générale, historique et archéologique, à travers la région occupée (2). Le P. Azais, chargé du service archéologique, a fait, à cette occasion, un dénombrement des sites archéologiques de la Thrace occidentale. A la suite de cette enquête, certaines fouilles ont été entreprises elles devaient s'étendre à tous les <i>tumuli</i> de la région de Gumuldjina. Deux seulement ont pu être explorés (fig. 1) ; les objets trouvés (fig. 2) révèlent qu'ils couvraient l'un et l'autre des sépultures assez tardives (époque romaine) (3).</p> <div style="text-align: center;">  </div> <p style="text-align: center;">Fig. 2 - Environs de Gumuldjina. Vase de verre (haut. 0,16) provenant d'un <i>tumulus</i>.</p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

FIGURE 1 EXEMPLE DE CONSULTATION D'UN DOCUMENT SEMI-STRUCTURÉ

## **CONCLUSION**

Au fil de notre étude, nous avons mis en évidence comment la représentation d'un document sous forme de données semi-structurées pouvait résoudre un certain nombre des questions posées par un projet de bibliothèque virtuelle abritant le corpus d'un périodique scientifique : le respect des particularités d'une telle œuvre, sa numérisation, et son utilisation. Nous avons donc posé la première pierre du projet, et en connaissons les principales directions à approfondir : étude du marché (logiciels et services) de la technologie XML/XSL, modélisation sémantique d'un périodique scientifique, dématérialisation d'un document scientifique multilingue, communications interpersonnelles dans la bibliothèque virtuelle.