Chapitre 10. CEFAEL : Collections de l'Ecole française d'Athènes en ligne*

Comme nous l'avons vu dans le chapitre précédent, l'École française d'Athènes publie une douzaine de collections (séries de monographies et revue). Depuis 1877, ce sont près de 570 volumes, soit 250.000 pages, qui ont été édités par l'Ecole. Si un tel corpus représente un intérêt indéniable pour l'historien des sciences, il n'en est pas moins important pour l'archéologue. En effet, la fouille archéologique présente la particularité de détruire les couches qu'elle étudie. Que reste-t-il alors de son objet d'étude ? Le carnet de fouille, l'article, la monographie… Ainsi, un article de 1877, malgré les révolutions théoriques et de méthodologiques qu'a pu connaître la discipline depuis, reste-t-il un substitut incontournable des vestiges qu'il décrit.

A l'heure où des fondations américaines mettent en place d'immenses bibliothèques numériques (comme JSTOR[87]) portant sur la rétrospective des revues en Sciences Humaines, le Ministère de la Recherche a souhaité encourager les expérimentations technologiques permettant à terme, au niveau français ou européen, des alternatives publiques. C'est ainsi qu'en décembre 2001 le projet de mise en ligne des collections de l'Ecole a reçu le soutien financier du « Plan de numérisation des publications en SHS ». Aujourd'hui, le portail CEFAEL[88] permet, à travers de multiples structures hypermédia, de feuilleter gratuitement sur la Toile l'intégralité des pages du corpus (sous forme de fac-similés).

La nature du projet nécessitait une valorisation immédiate du corpus à l'aide de technologies éprouvées. Cependant, comme nous allons le voir dans ce chapitre, il a été possible d'expérimenter la gestion avec *Porphyre* d'une partie de ce corpus. Dans une première partie, nous étudierons qu'elles sont les différentes structures hypermédia dont a besoin le lecteur. Dans une deuxième partie, nous présenterons la chaîne de numérisa-

---

* Des parties de ce chapitre ont fait l'objet d'une conférence lors de la journée d'étude sur les bibliothèques numériques [Benel02b].

87 http://www.jstor.org

88 http://cefael.efa.gr

tion et de diffusion mise en œuvre dans le projet CEFAEL. Ensuite, dans une troisième partie, nous exposerons le protocole expérimental proprement dit. Enfin dans une quatrième partie, nous verrons les problèmes rencontrés et les solutions proposées.

1. Etude des besoins

Afin de définir les différentes structures hypermédia nécessaires à l'interprétation des collections de l'Ecole, nous allons tenter d'identifier les « points de vue » de différents acteurs intervenant sur une même page du corpus (cf. Figure 10.1).

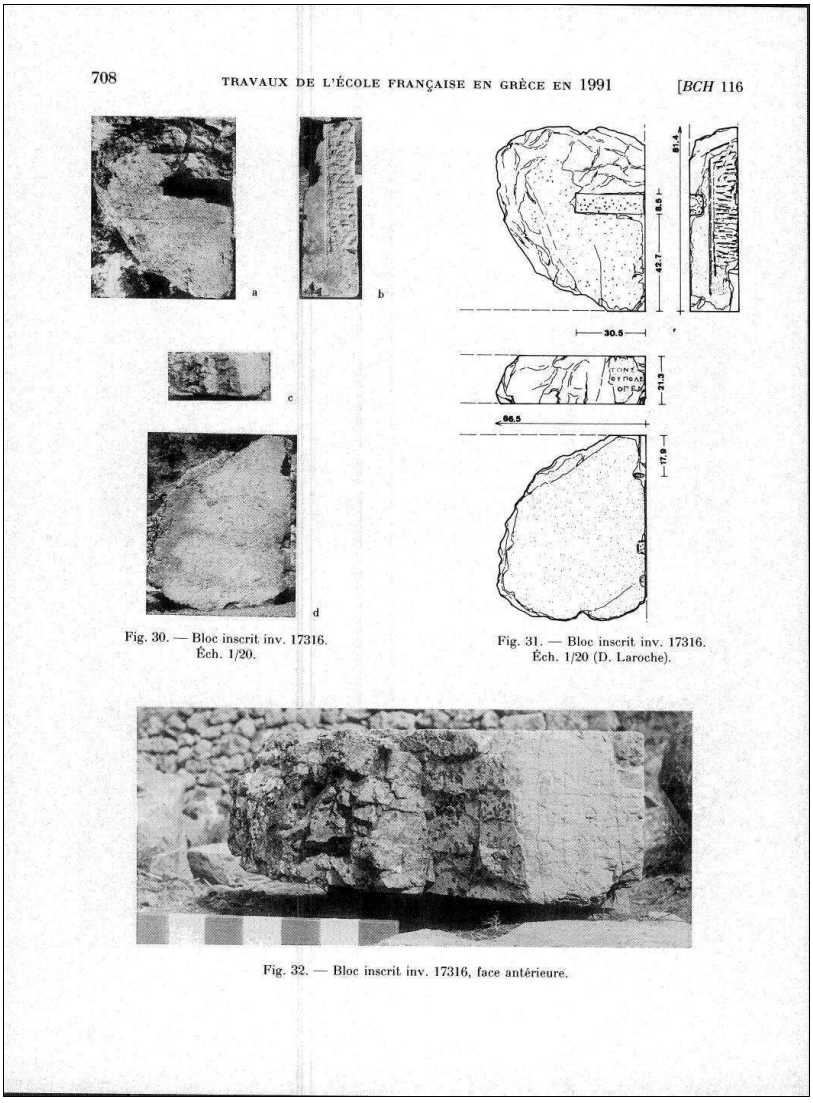

Figure 10.1 : Une page à étudier sous différents points de vue.

a. Maquettiste

La première structuration du corpus est donnée par le maquettiste : il s'agit de la pagination. Chaque page est ainsi désignée sans ambiguïté par le triplet « Collection/Volume/Folio ». Cette nomenclature arborescente permet ainsi de nommer la page choisie « BCH/116/708 » (cf. Figure 10.2). On peut utiliser des folios spéciaux pour ceux habituellement en chiffre romain (pages préliminaires) et pour les pages non foliotées (pages finales, dépliants, planches…). Notons qu'il n'est pas indispensable d'introduire le niveau du tome puisque la pagination est continue d'un tome au suivant.

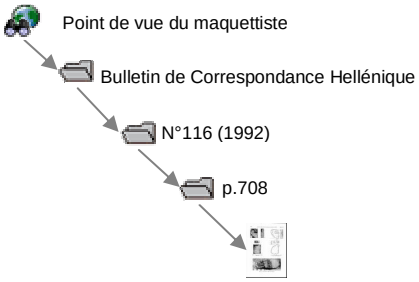

Figure 10.2 : Extrait de la facette du maquettiste (Réseau de description *Porphyre*)

b. Bibliothèque

Si la structure précédente suffit à référencer l'ensemble du corpus, le chercheur a cependant besoin d'autres structures pour y accéder. L'une de ces structures est celle qui apparaît dans le catalogue de la bibliothèque. Cette structure identifie au sein des volumes des éléments que l'on appellera « publications » (articles de recherche, rapports, chroniques…). Ces publications ont pour attribut une date et un ou plusieurs auteurs. Notre page d'exemple (cf. Figure 10.3) appartient à un rapport sur les travaux de l'Ecole à Delphes, daté de 1991, et cosigné par les huit auteurs indiqués.

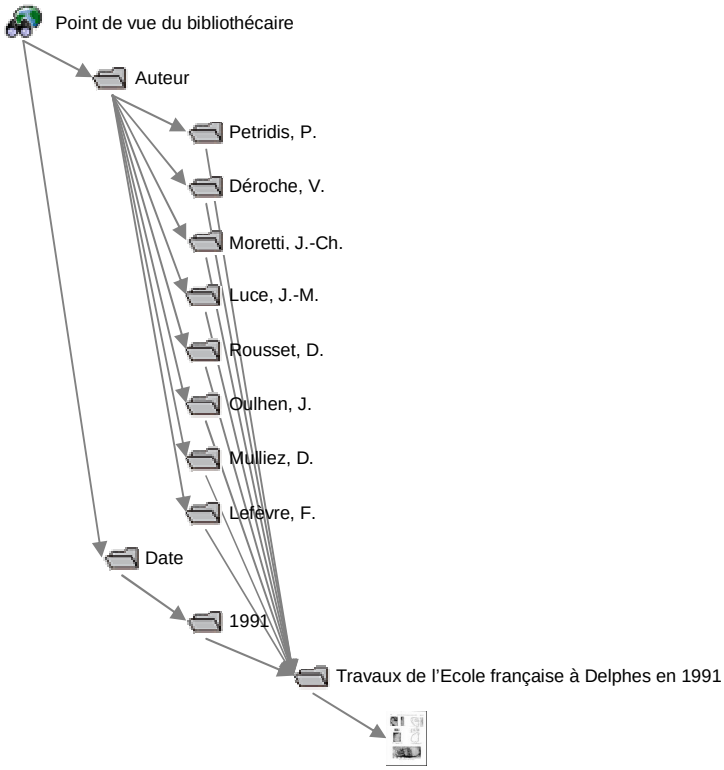

Figure 10.3 : Extrait de la facette du bibliothécaire (Réseau de description *Porphyre*)

c. Photothèque/Planothèque

L'Ecole dispose d'un fond de près de 500.000 photographies et plans datant de la fin du XIX siècle à nos jours. Ce fond comprenant entre autres les figures publiées dans les collections, on peut considérer que la structure du fond est aussi structure de la collection. Ainsi, la Figure 10.4 montre-t-elle que notre page d'exemple comprend deux figures correspondant aux photographies d'archive « R3879-007 » et « L9689-030 ». Chacune de ces photographies peut être décrite par un certain nombre de « méta-données », telles que leur auteur (Jean-Charles Moretti) et leur date de prise de vue (1991).

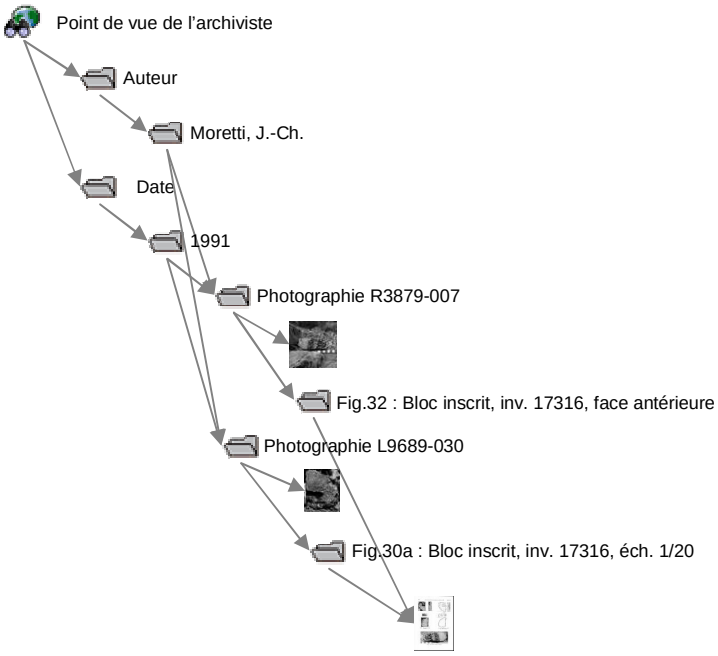

Figure 10.4 : Extrait de la facette de l'archiviste (Réseau de description *Porphyre*)

d. Equipe de fouille

Les trois premières structures étudiées sont loin d'être exhaustives. En effet le corpus est appelé à être structuré par chacun de ses lecteurs. Un exemple intéressant nous est donné par l'équipe de fouille de Roland Etienne. Cette équipe travaille actuellement à analyser la bibliographie concernant le sanctuaire de Délos en fonction de la position spatiale de chacun des vestiges décrits. La plupart de cette bibliographie étant contenue dans les collections de l'Ecole, on peut donc considérer que l'on est en présence d'une nouvelle structure du corpus. De la même manière, notre page d'exemple pourrait entrer dans une structure correspondant à la géographie du site de Delphes (cf. Figure 10.5).

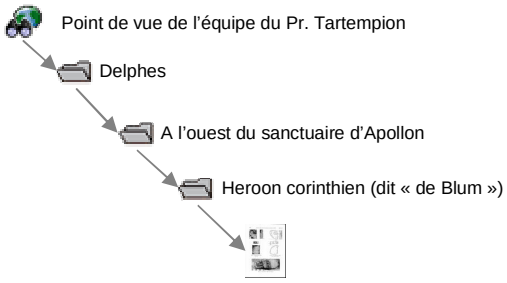

Figure 10.5 : Extrait de la facette d'une équipe de fouille (Réseau de description *Porphyre*)

2. Numérisation et valorisation

La numérisation du corpus démarra en mai 2001, à Lyon, sur le site de la plate-forme technologique du CNRS à la Maison de l'Orient et de la Méditerranée [MOM]. Les corpus complets, disponibles en bibliothèques, ne pouvant être ravis aux lecteurs, il fut nécessaire, pour reconstituer les collections, de faire venir des volumes d'Athènes (EFA, éditeur), de Limoges (Bontemps, imprimeur) et de Paris (De Boccard, distributeur).

Chaque volume fut, préalablement à la numérisation, décrit dans une base de données (nombre de pages foliotées, nombre de planches, etc.) et massicoté. Cette dernière opération permit une numérisation de masse[89] utilisant un scanner recto-verso à chargeur[90].

A la sortie du scanner, nous disposions de répertoires contenant des images à haute définition compressées sans pertes[91], numérotées automatiquement. En se basant sur la description des volumes, nous pûmes automatiquement[92] produire :

- un rapport permettant de contrôler que le nombre de pages numérisées était cohérent avec la description des volumes,
- des archives, sur différents supports[93], où chaque image brute était renommée en fonction des méta-données du volume,

[89] Les dépliants, quant à eux, durent être numérisés « à la main ».

[90] Xerox Digipath.

[91] TIFF, compression CCITT Group 4, 600 points par pouce.

[92] Grâce au « Robot Transvision », logiciel développé à la MOM.

- des images pour la diffusion sur le Web (à une définition inférieure et compressées avec pertes[94]) ainsi que des vignettes.

La diffusion sur la Toile [Benel02b] est rendue possible à l'aide de deux types de serveurs HTTP. Le premier[95] permet de stocker les fac-similés et de les redimensionner en fonction des besoins de l'usager (taille de son écran). Tandis que le second[96] génère l'hypertexte permettant de feuilleter ces fac-similés. Notons que contrairement à ce que permettrait *Porphyre,* cet hypertexte n'autorise la navigation que dans une facette à la fois.

CEFAEL est hébergé au CINES, et profite donc de la puissance des machines du centre, de son réseau très haut-débit (nœud régional RENATER), et surtout de son équipe disponible 24h/24, 7j/7.

Pour conclure cette section, notons que la chaîne de production ainsi décrite permit d'atteindre, avec deux personnes affectées à la description et à la numérisation des ouvrages, une productivité de 40.000 pages par mois [Iacovella 2002].

3. Expérimentation dans Porphyre

Notre expérimentation eut lieu au cours de l'été 2001, au moment où seul un petit corpus de test avait été numérisé et était disponible dans l'intranet de l'EFA. La description du corpus fut exportée de la base de données vers *Porphyre.* Pour ce faire, nous dûmes définir un format d'échange pour les réseaux de description (à l'aide d'une DTD[97]), réaliser une petite « moulinette » pour générer le fichier correspondant à la base de données du corpus, ainsi qu'ajouter à *Porphyre* un module d'import pour ce type de fichiers.

---

[93] Notons tout de même que le nombre de CD-ROMs nécessaires à l'archivage du corpus est de l'ordre de 250 !

[94] JPEG, niveaux de gris, 150 points par pouce.

[95] Utilisant le système Transvision® développé par la MOM.

[96] Serveur « web » (Apache) agrémenté de scripts (développés en PHP) et d'une base de donnée (Sybase).

[97] Définition de type de document XML.

L'exploitation dans *Porphyre* de ce corpus de test fit l'objet d'une démonstration (cf. Figure 10.6) aux Journées Bibliothèques Numériques de mai 2002. Par ailleurs, en important un grand nombre de fois les mêmes volumes, nous pûmes tester la montée en charge des serveurs.

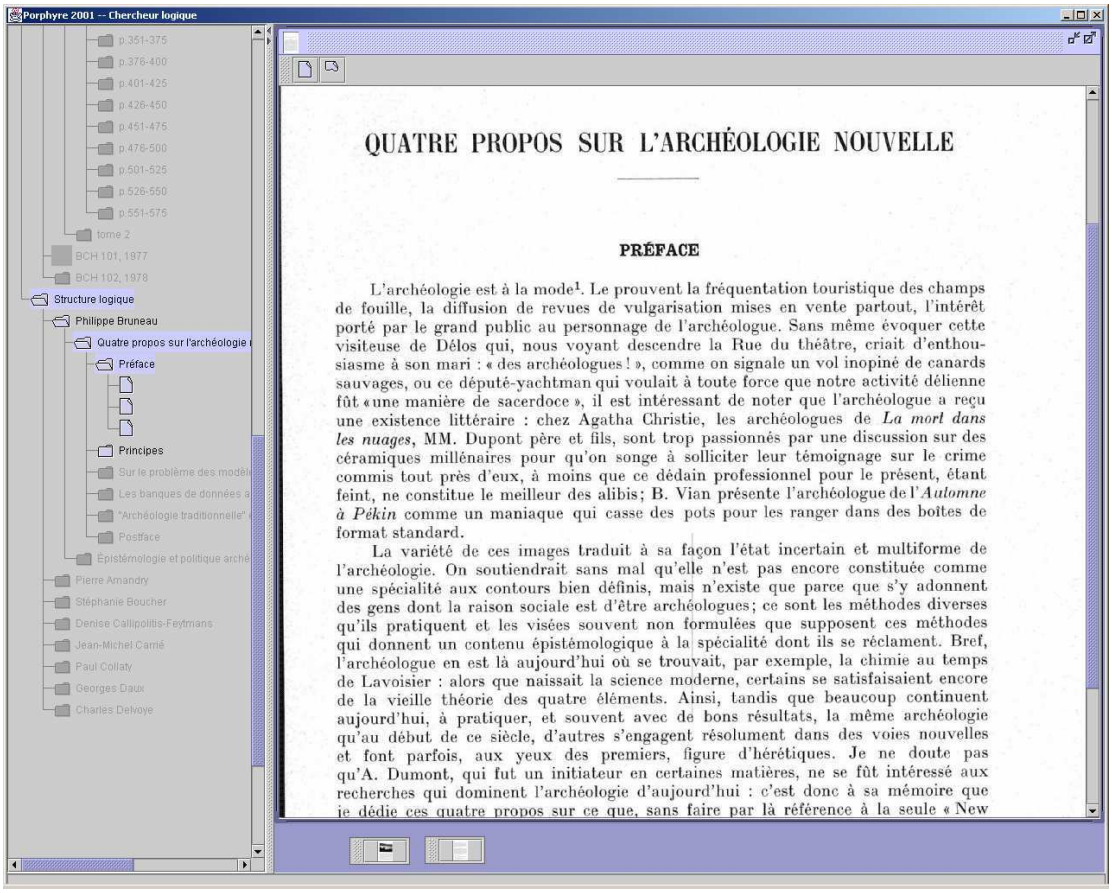

Figure 10.6 : Lecture avec *Porphyre 2001* d'un extrait des *Collections de l'École française d'Athènes en ligne*

4. Retour d'expérience

Le prototype d'alors était basé sur un serveur de contenu et un serveur de structure. Le premier était constitué de « servlets » appliquant des algorithmes « maison » à des images JPEG. Le second était conçu de telle sorte que les réseaux de description puissent dépendre les uns des autres suivant un ordre partiel. Par exemple, tout descripteur du réseau d'une bibliothèque pouvait être généralisé par un descripteur du réseau

d'un chercheur, à condition que ce chercheur soit « abonné » à la bibliothèque. Chaque serveur devenait alors le client de plusieurs autres.

Le premier problème rencontré concernait le serveur de contenu. D'une part, il était regrettable de ne pouvoir gérer que des versions dégradées (JPEG) des fac-similés. Ensuite, la performance de l'architecture à base de servlets et d'algorithmes « maisons » s'est avérée insuffisante. La nouvelle version à base de scripts PHP, intégrant des composants externes optimisés, a permis un gain de performance considérable [Tribollet03].

Le second problème concernait le serveur de structure. Le mode de distribution des données ne permettait de tirer aucun profit de la mise en parallèle des calculs sur les différents serveurs. Pour remédier à cela, nous avons défini les notions d'objets documentaires et de facettes. Aujourd'hui deux réseaux de description ne dépendent l'un de l'autre que par l'intermédiaire des objets documentaires. Au niveau de l'architecture, le client interroge directement les serveurs. L'intégration des données est rendue possible par le fait que les serveurs se réfèrent aux mêmes serveurs de correspondance. Avec la nouvelle architecture, si l'on gère *n* facettes sur *n* serveurs différents la charge des serveurs sera *n* fois moindre que sur un serveur unique. Une autre optimisation est également envisagée. Elle consisterait à tirer partie du fait que la facette du maquettiste est arborescente. Le filtre étant beaucoup moins complexe à calculer avec de telles structures, il serait judicieux de développer un serveur spécialisé implémentant le même protocole mais de manière optimisée.