Du fac-similé à l'hypertexte

Aurélien Bénel

Du fac-similé à l'hypertexte

Aurélien Bénel
Aurelien.Benel@lisi.insa-lyon.fr

I. Introduction

Numériser des collections, pour quoi faire ? Beaucoup s'interrogent sur l'intérêt de la numérisation sous forme de fac-similés (i.e. en mode image) et se demandent en quoi il s'agirait d'un plus pour la consultation savante. Ce type de consultation est en effet caractérisé par une localisation par « coup d'œil » de l'information cherchée suivie d'une lecture approfondie des fragments choisis. Or, on peut se demander si l'on pourra *feuilleter* les ouvrages d'un corpus numérisé comme on feuillette ceux qui sont sur l'étagère ?

Pour répondre, nous devons être conscients que la pratique consistant à feuilleter des pages (« browsing » en anglais) est au centre même de la notion « d'hypertexte ». Cependant, faudrait-il encore trouver un procédé permettant de transformer les fac-similés des pages en un véritable hypertexte…

Un certain nombre de travaux de recherche prometteurs portent sur la « dématérialisation » (ou rétro-conversion) des fac-similés, c'est à dire le passage d'une image à une page « web » en texte intégral. Cependant l'état actuel de ces techniques ne nous permet pas de les intégrer dans la perspective opératoire qui est la nôtre.

L'objet de ce chapitre est de proposer une alternative simple et efficace. Tout d'abord, nous essaierons de préciser la notion d'hypertexte comme étant le couple contenu-structures. Ensuite, nous essaierons de voir suivant les types de structure comment les stocker et faciliter leur acquisition.

II. Hypertexte : Contenu et Structures

1. Retour aux origines

Chacun de nous a tendance à associer la notion « d'hypertexte » au « World Wide Web », c'est-à-dire à une version simplifiée proposée en 1989-90 par le CERN[1] pour ses besoins propres. Or, il faut faire remonter la notion aux années 1945[2] (même si le mot date de 1965[3]). Il s'agissait à l'époque de construire une machine (mécanique) permettant de « feuilleter » des microfilms ! Le principe était en effet le suivant : associer à un *contenu* documentaire interprétable uniquement par l'homme, une *structure* (ou parcours) gérable par une machine. Ainsi, la notion d'hypertexte n'est pas antinomique avec celle de fac-similé. Au contraire, il s'agissait dès l'origine de la superposition d'une ou plusieurs structures à un corpus de fac-similés.

2. Une architecture opératoire… dès maintenant !

L'obtention d'un hypertexte basé sur des fac-similés a été expérimentée dans le cadre des *Collections de l'Ecole Française d'Athènes En Ligne* (CEFAEL). L'architecture retenue repose sur les standards du « Web », c'est-à-dire sur le protocole HTTP et les formats HTML et JPEG, permettant ainsi à n'importe quel « butineur » (Netscape, Mozilla, Internet Explorer…) de consulter le corpus.

---

[1] TIM BERNERS-LEE, *Information Management: A proposal,* Rapport interne, CERN, Mars 1989.
[2] VANNEVAR BUSH, « As we may think », in *The Atlantic Monthly*, Juillet 1945.
[3] THEODOR H. NELSON, « The Hypertext », in *Proceedings of the World Documentation Federation*, 1965.

En ce qui concerne les fac-similés, ceux-ci sont stockés sur un ordinateur (éventuellement plusieurs) appelé « serveur de contenu ». Les parcours hypertextuels sont quant à eux stockés sur un ordinateur appelé « serveur de structure ». Le premier redimensionne les images en fonction des besoins de l'usager (taille de son écran). Tandis que le second génère l'hypertexte contenant ces images.

Dans notre prototype (cf. Fig.1), le premier utilise le système *Transvision* (développé par la Maison de l'Orient Méditerranéen) ; le second, un serveur « web » (*Apache*) agrémenté de scripts (développés en *PHP*) et d'une base de donnée (*PostgreSQL*).

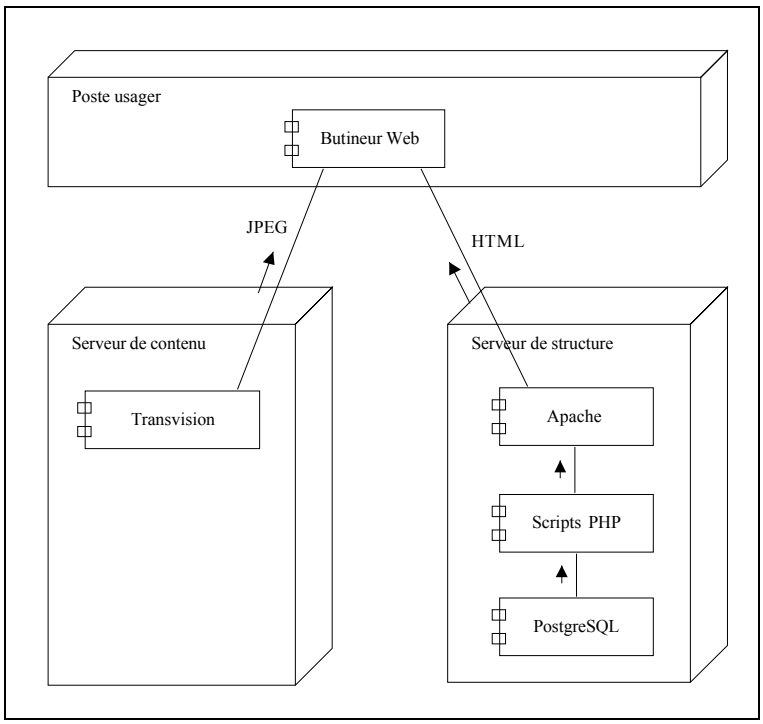

Figure 1. Architecture du prototype CEFAEL

Les questions d'architecture étant résolues, on peut maintenant s'interroger sur les différents types de structure à traiter. Dans le domaine de la « documentique », il a souvent été fait appel à une typologie distinguant les structures physique, logique et sémantique[4]. Pour un certain nombre de raisons, nous préférons (au sein du groupe « documents multi-structurés » de l'ISDN) adopter une typologie basée sur les usages.

Dans le cas qui nous intéresse de collections scientifiques, nous pourrions retenir la typologie suivante :

- structures éditoriales,
- structures « auctoriales »,
- structures « lectoriales ».

Dans la suite de notre propos, nous détaillerons ce que nous entendons par chacun de ces types de structure et comment nous allons les gérer.

[4] MARC NANARD, JOCELYNE NANARD, JACQUES CHAUCHE, ANNE-MARIE MASSOTTE, ALAIN JOUBERT et HENRI BETAILLE, « La métaphore du généraliste : Acquisition et utilisation de connaissances macroscopiques sur une base de documents techniques », in *Acquisition et ingénierie des connaissances,* Cépaduès Editions, 1996.

III. Structure éditoriale première

1. Principe

Par « structure éditoriale première », nous entendons uns structure donnée par l'éditeur permettant de désigner sans ambiguïté les éléments atomiques d'un corpus (ici : les pages). Ainsi une référence bibliographique comme « *BCH 1, tome 1,* p.1 », ou la nomenclature correspondante `BCH_1_1_T_1`, identifie, sans ambiguïté aucune, la première page textuelle du premier tome du premier numéro du BCH.

La nomenclature des pages dessine en fait une structure arborescente (cf. Fig.2) sur quatre niveaux :

- `BCH` correspond à la **Série**
- `BCH_1` au **Numéro**
- `BCH_1_1` au **Volume** (ou tome)
- `BCH_1_1_T1` à la **Page**

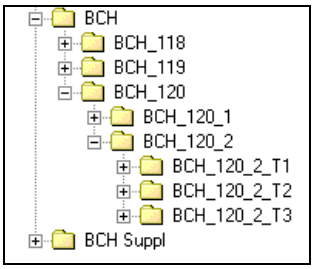

Figure 2. Structure arborescente de la nomenclature des pages

Dans la base de donnée CEFAEL (cf. Fig.3), chacun de ses niveaux correspond à une table possédant ses attributs propres (ex : année de publication pour chaque numéro, URL pour chaque page…).

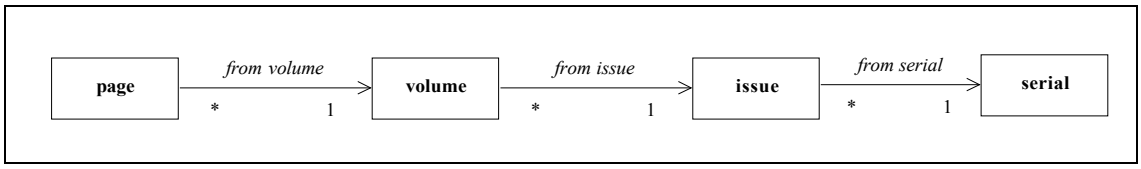

Figure 3. Extrait du schéma de la base de donnée CEFAEL.

2. Génération

Dans les articles précédents, il a été présenté comment à partir de la description du volume, chaque page numérisée était associée dans *Transvision* à sa nomenclature. Comme *Transvision* fait serveur « web », il associe également à chaque page une URL permettant d'y accéder. La génération dans CEFAEL de la structure éditoriale première consiste donc à exporter, à partir de *Transvision,* les couples nomenclature-URL et à les importer dans la base de donnée CEFAEL.

Actuellement cette génération se fait à l'aide d'un certain nombre de « moulinettes » (EFAexport, MrSplit, requêtes SQL) suivi de la saisie de quelques informations complémentaires (cf. Fig.4). D'ici peu, l'ensemble sera piloté par le robot et les informations complémentaires seront tirées de la description des volumes (de l'étape de numérisation).

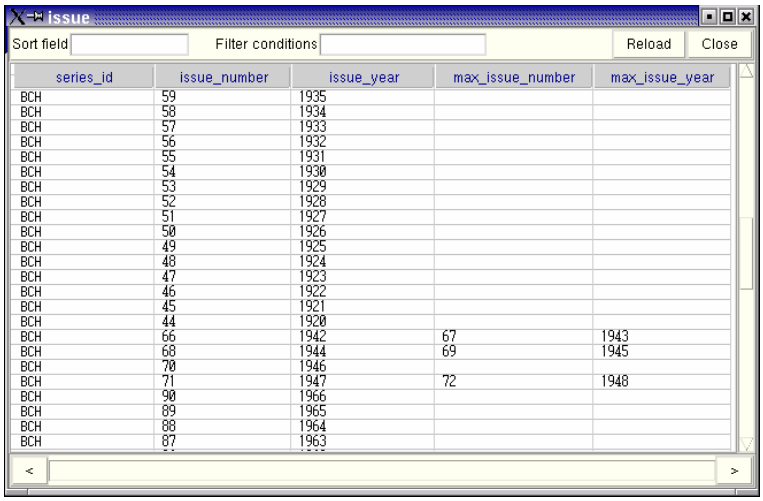

Figure 4. Saisie d'informations complémentaires (années, numéros multiples pour les années de guerre).

En ce qui concerne les pages « web », il n'y a rien de plus à faire. Elles seront générées dynamiquement en fonction des requêtes des usagers (grâce aux scripts PHP).

3. Consultation

Interrogation

L'interrogation de la *structure éditoriale première* se fait en fonction de la série, du numéro, de l'année et du tome (cf. Fig.5). A défaut de critères spécifiés par l'usager, c'est la liste de l'ensemble des volumes du corpus qui est demandée.

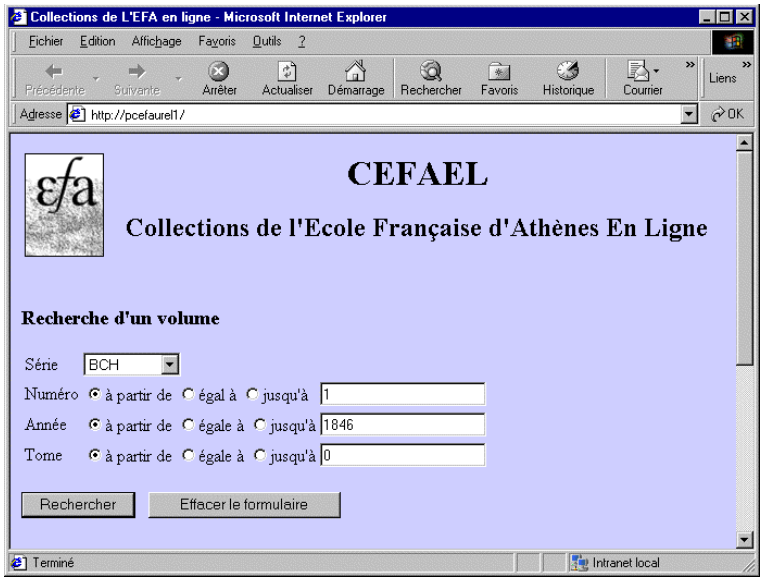

Figure 5. CEFAEL : Ecran d'interrogation portant sur les volumes.

Sélection d'un volume

L'ensemble des volumes correspondant à la requête est affiché sous forme de liste de références bibliographiques (cf. Fig.6). La sélection d'un volume se fait en cliquant sur l'icone[5] correspondant.

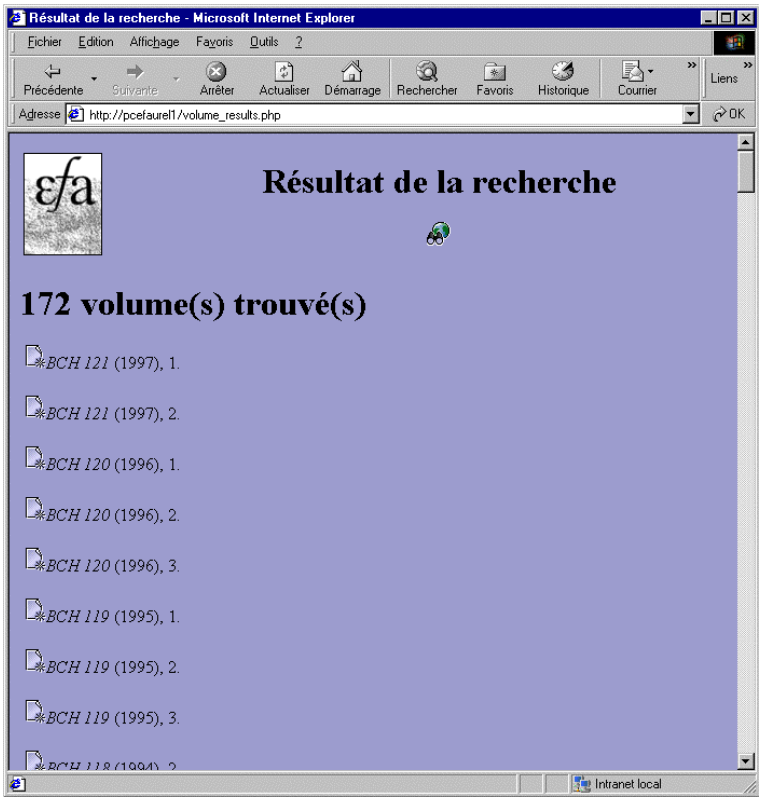

Figure 6. CEFAEL : Ecran de sélection d'un volume.

Affichage du fac-similé

La sélection d'un volume provoque l'affichage du fac-similé. Est affichée la première page indiquée comme table des matières. Il est possible de se déplacer vers la page suivante et précédente, de revenir à la table des matières, et de choisir dans la « liste déroulante » n'importe quelle page du volume.

Etant donné que les fac-similés archivés sont d'une précision bien supérieure à l'usage courant, les fac-similés sont, par défaut, réduits[6] à la largeur de l'écran. Il est cependant possible de choisir d'afficher plus ou moins de détails à l'aide des deux boutons de zoom.

[5] Après vérification dans le Robert, l'icone informatique est bien loin de ses origines byzantines : il a perdu sa féminité et son accent !

[6] La réduction de la taille des fac-similés est effectuée de manière dynamique par le serveur de contenu.

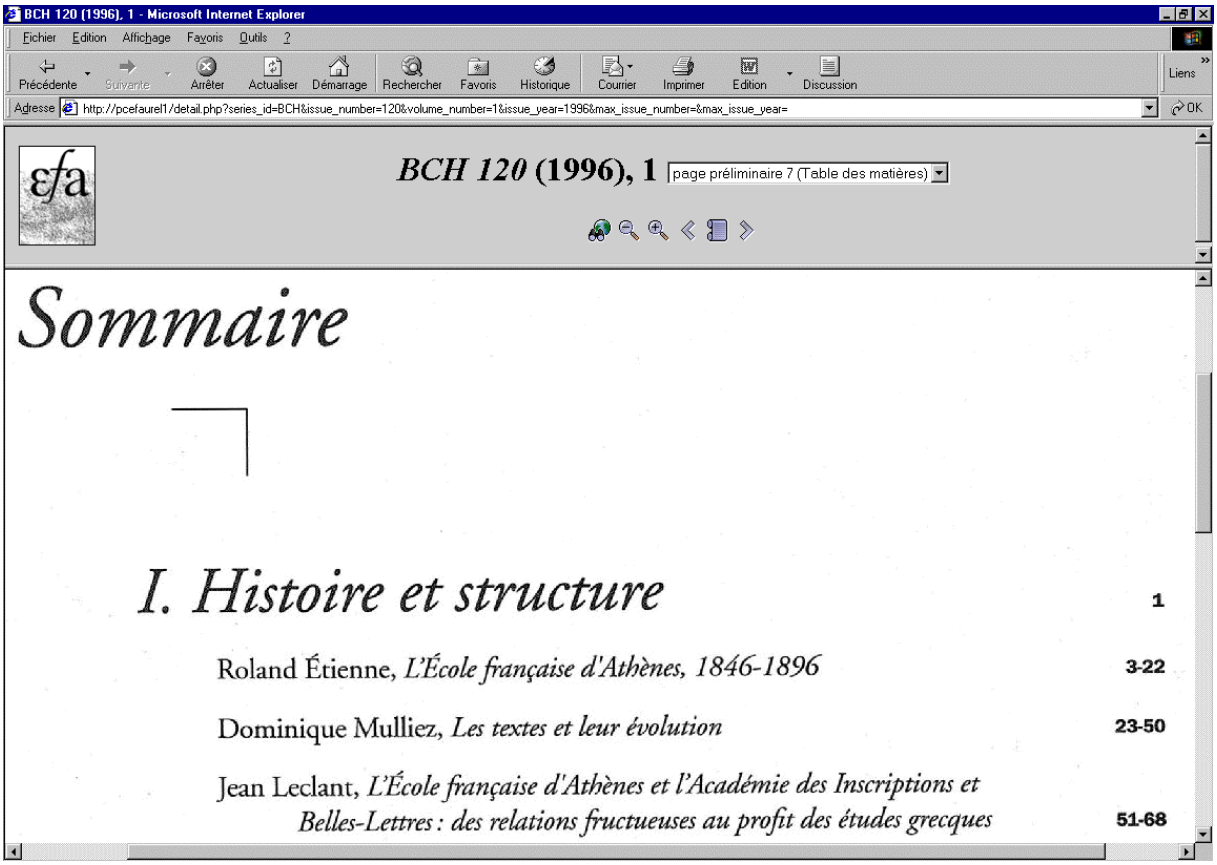

Figure 7. CEFAEL : Ecran de lecture d'un volume.

IV. Structure éditoriale complémentaire

1. Principe

Si la structure éditoriale primaire suffit à référencer l'ensemble du corpus, le chercheur a cependant besoin d'autres structures d'accès au corpus. L'une de ses structures apparaît dans le sommaire : il s'agit de la « structure éditoriale complémentaire ». Celle-ci va permettre d'identifier au sein des volumes des éléments que l'on appellera « publications » (articles de recherche, rapports, chroniques…).

Comme l'indique le schéma de la base de donnée (cf. Fig.8), une publication peut être signée par plusieurs auteurs (et un auteur signer plusieurs publications). Elle est matérialisée par une séquence de pages désignée par une page de début et une page de fin. Il est également possible, lorsque l'on dispose d'une version en texte intégral, de faire référence à cette dernière.

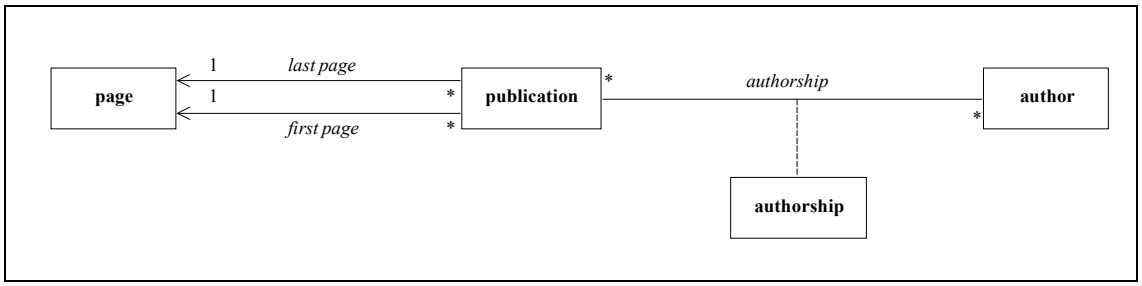

Figure 8. Extrait du schéma de la base de donnée CEFAEL.

2. *Saisie*

Compte tenu du faible volume d'information que représente la structure éditoriale complémentaire, une saisie de cette structure paraît tout à fait raisonnable. Elle est réalisée à travers des pages « web » accessibles en interne (« intranet »). Ces pages permettent de mettre à jour, d'une part, les informations portant sur les publications (cf. Fig.9) et, d'autre part, la liste d'autorité des auteurs (cf. Fig.10).

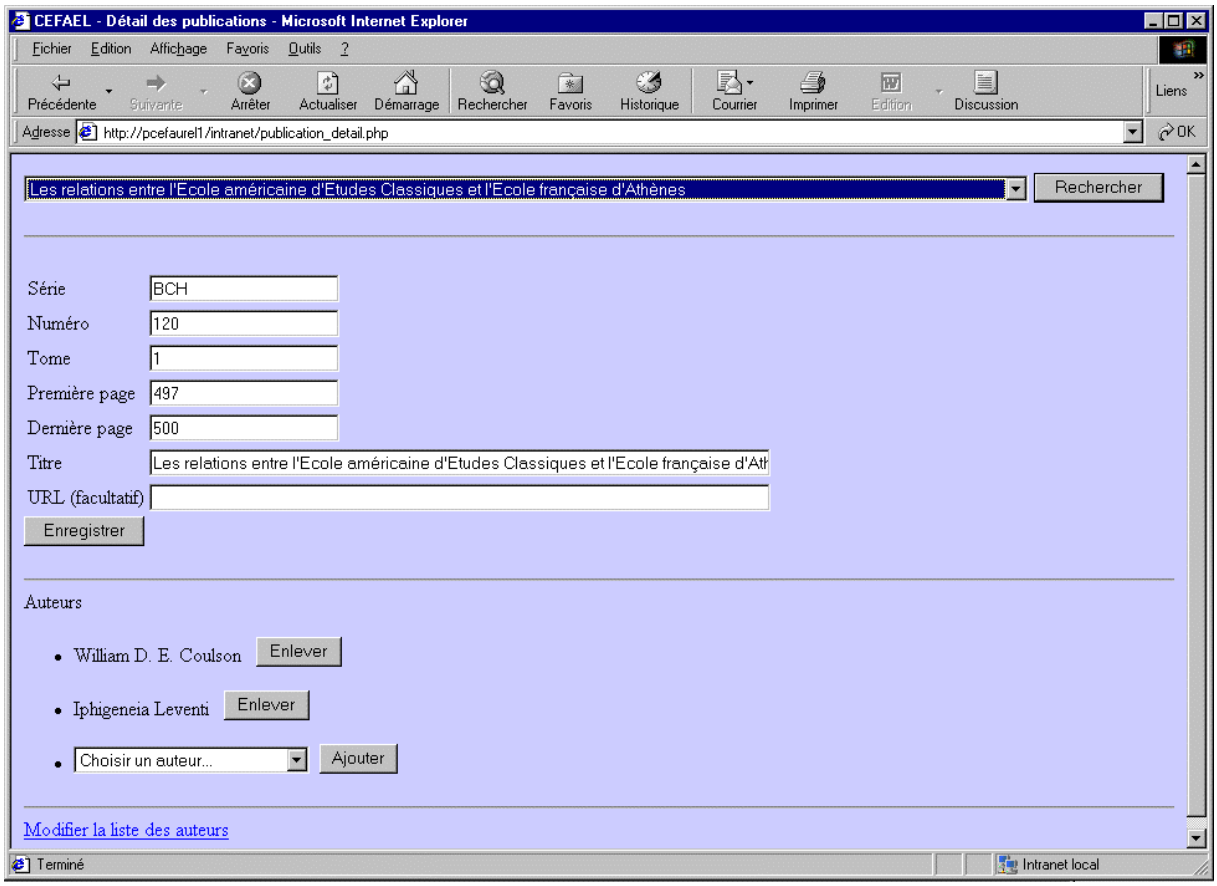

Figure 9. CEFAEL (Intranet) : Saisie des informations portant sur les publications.

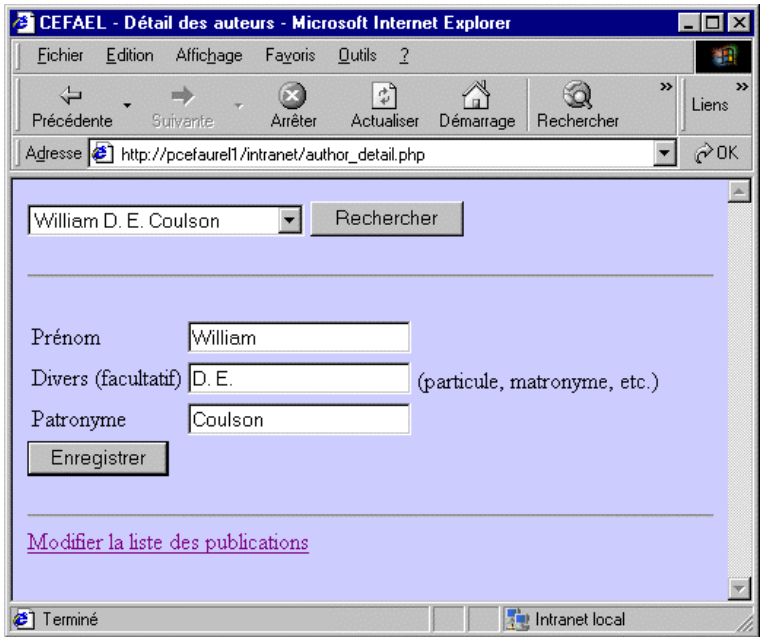

Figure 10. CEFAEL (Intranet) : Saisie des autorités « auteurs ».

3. *Consultation*

Interrogation

L'interrogation de la structure éditoriale complémentaire se fait en fonction des auteurs et de mots du titre (cf. Fig.11). Il est également possible de croiser ces critères avec ceux de la structure éditoriale première (série, numéro, année et tome). A défaut de critères spécifiés par l'usager, c'est la liste de l'ensemble des publications du corpus qui est demandé.

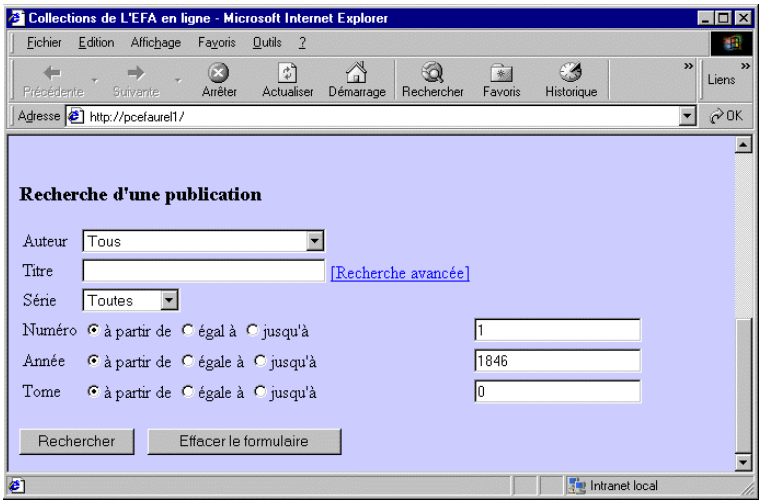

Figure 11. CEFAEL : Ecran d'interrogation portant sur les publications.

Sélection d'un article

L'ensemble des articles correspondant à la requête est affiché sous forme de liste de références bibliographiques (cf. Fig.12). La sélection d'un article se fait en cliquant sur l'icone correspondant.

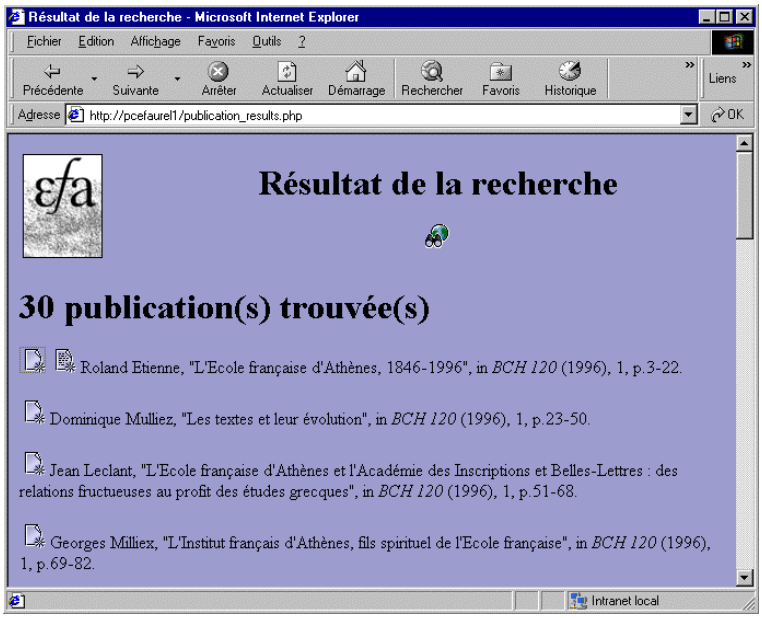

Figure 12. CEFAEL : Ecran de sélection d'une publication.

Affichage du fac-similé

L'interface de consultation d'un article est identique à celle d'un volume à la différence près qu'est affichée par défaut la première page de l'article (au lieu de la première page de la table des matières) et que sont distinguées, dans la « liste déroulante », les pages appartenant à l'article choisi, des autres pages du volume.

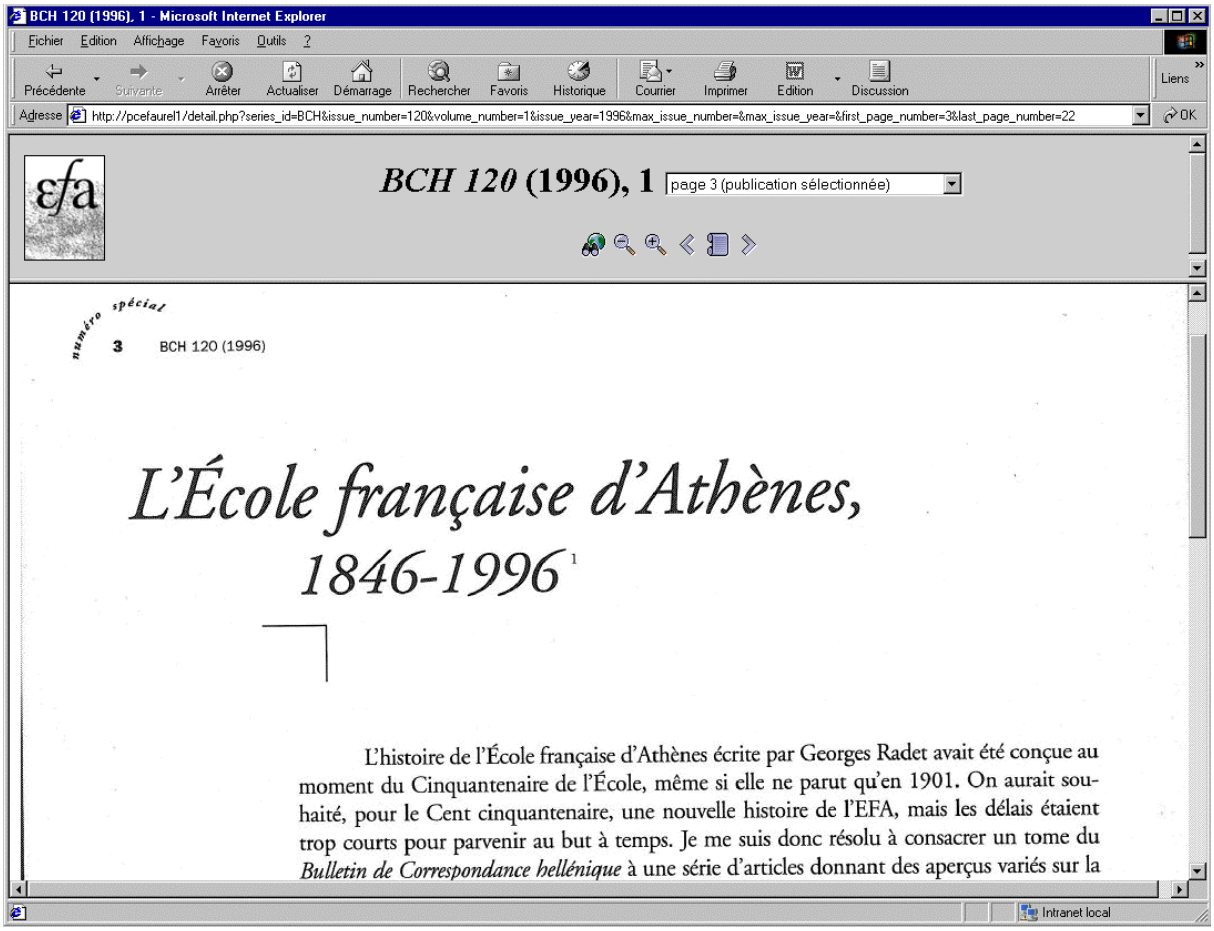

Figure 13. CEFAEL : Ecran de lecture d'une publication.

V. Perspectives

Ici s'achève la présentation de l'état actuel des développements en ce qui concerne les *Collections de l'EFA En Ligne*. Cependant, fort d'autres expérimentations, nous pouvons donner un aperçu de ce à quoi nous visons à moyen terme.

Les premières limites du modèle proposé sont qu'il existe différentes normes de description d'une structure éditoriale (Dublin Core, UNIMARC…) et que se limiter à une seule entraverait la capacité d'interconnexion de la bibliothèque numérique avec d'autres systèmes documentaires (catalogues ou corpus en ligne).

D'autre part, il faudrait s'intéresser à d'autres structures du corpus, à savoir : la structure « auctoriale » et les structures « lectoriales ».

1. Structure « auctoriale »

Par structure « auctoriale », nous appelons la structure que l'auteur a souhaité faire ressortir par la forme de sa publication (typographie, mise en page, etc.). L'exemple le plus courant est

celui de la hiérarchie des sections (cf. Fig.14). Dans ce cas là, il est important de noter que ni la profondeur, ni la signification des niveaux ne sont donnés par un modèle *a priori.*

Rajoutons que la structure « auctoriale » n'est pas forcément une simple hiérarchie, la *Chronique des fouilles* du BCH présente par exemple une structure à plusieurs facettes (ou dimensions) : par site de fouille, par année et par équipe (cf. Fig.15).

Dans tous les cas, nous avons affaire à ce que les informaticiens appellent des « modèles semi-structurés », c'est à dire des modèles qui ne peuvent être stockés tels quels dans des bases de donnée classiques[7].

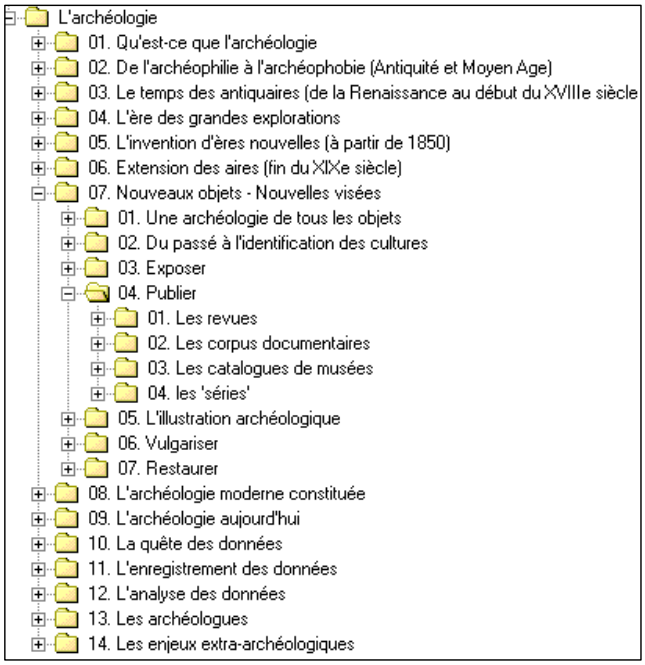

Figure 14. Exemple de structure « auctoriale » hiérarchique.

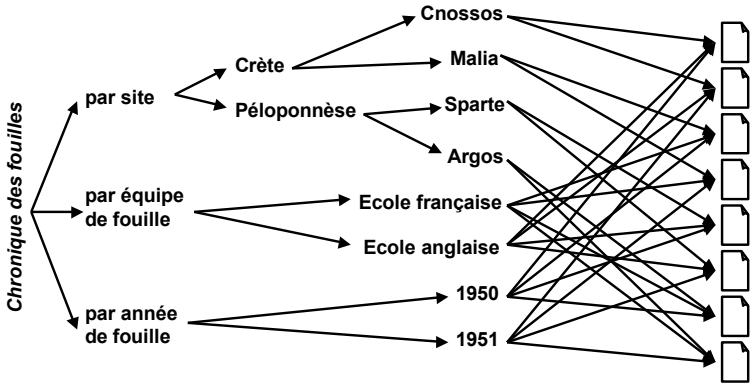

Figure 15. Exemple de structure « auctoriale » à plusieurs facettes.

2. Structures « lectoriales »

Les structures les plus intéressantes sont sans doute les structures « lectoriales ». Il s'agit en effet de considérer les traces d'interprétation du lecteur expert :

- la grille de lecture à travers laquelle il va étudier le texte,
- l'appareil critique qu'il va choisir d'utiliser,

[7] A moins d'effectuer une « méta-modélisation », c'est-à-dire une modélisation du modèle…

- les relations qu'il va tisser avec d'autres textes par le biais de sa bibliographie,
- les notes de lecture qu'il va prendre,
- les articles, monographies et autres documents qu'il va écrire…

Ainsi chaque structure construite par le lecteur peut devenir un nouveau point d'entrée dans le corpus. Cependant, ceci pose un certain nombre de difficultés. Tout d'abord, nous devons définir différents espaces de communication : des espaces d'élaboration – privés ou communautaires (par exemple pour un directeur de recherche et ses doctorants), ainsi que des espaces de publication – permettant de « rendre public ». Autre difficulté, nous n'avons pas de *données*, mais des *hypothèses* dynamiques, exploratoires et non-consensuelles : signes du débat qui caractérise toute science. En fait, face à de telles attentes, les bases de données classiques apparaissent comme assez mal adaptées. Il devient nécessaire de penser à des outils spécialisés pour les chercheurs en Sciences Humaines.

3. *Vers un outil unifié*

Notre projet de recherche mené depuis 1998 consiste à imaginer et expérimenter un outil qui permettrait de gérer un corpus documentaire à travers les structures éditoriales, « auctoriales » et « lectoriales ». Ceci a abouti au prototype *Porphyre*[8], un système informatique permettant la manipulation conjointe de différentes structures de corpus, sans leur imposer un modèle *a priori*.

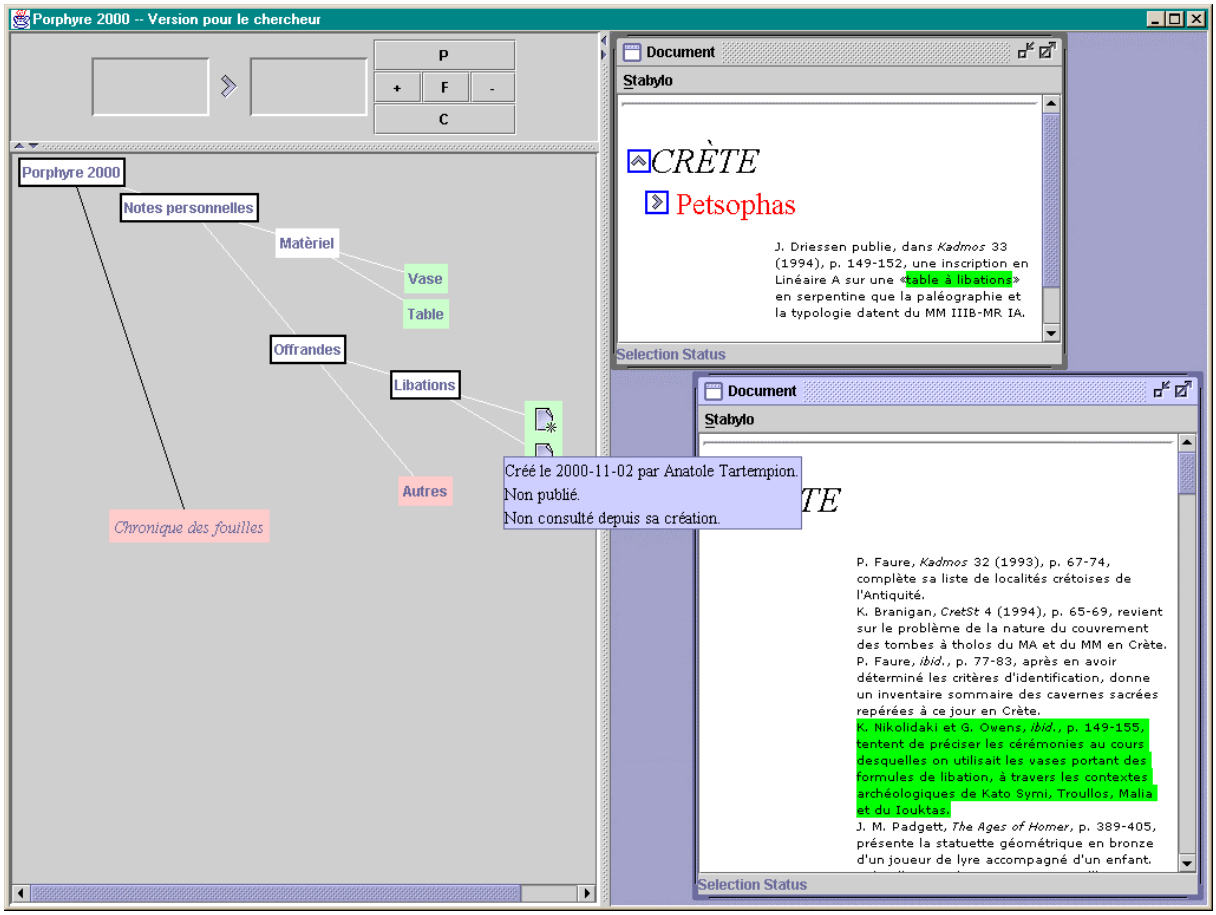

Figure 16. Porphyre : feuilletage d'un corpus à travers ses différentes structures.

[8] Une présentation du projet ainsi que les principaux articles de vulgarisation et de recherche pourront être trouvés à l'URL suivante : http://www.efa.gr/Informatique/porphyry.htm

L'ensemble des structures y est restitué, par auteur, sous la forme d'un graphe visuel qui permet le parcours de tous les documents à travers leurs différentes dimensions (cf. Fig.16). L'accès aux serveurs se décline en une consultation « web » ouverte à tous les publics et une consultation destinée aux experts, membres d'une communauté, pour publier des documents, les annoter et partager ces annotations. Le but d'un tel système, par les différents parcours qu'il offre à travers les annotations personnelles et étrangères, est d'assister le chercheur dans sa compréhension du corpus étudié.

VI. Conclusion

Ce chapitre s'inscrivait dans l'objectif général de proposer une chaîne de production visant à transformer une collection scientifique sur papier en un outil de travail électronique. Notre contribution a porté sur les moyens de « tisser » au-dessus de fac-similés numériques un certain nombre de structures, afin d'obtenir un hypertexte.

Nous avons étudiés dans le détail comment, dans le cadre du projet CEFAEL, nous avons pu générer dynamiquement un hypertexte en se basant sur la structure éditoriale stockée dans une base de données.

Si le besoin s'est esquissé de gérer d'autres structures comme les structures « auctoriales » et « lectoriales », nous avons vu en quoi leur complexité nécessitait d'utiliser des outils (comme le prototype *Porphyre*) plus adaptés que les bases de données aux Sciences Humaines.