Documents à structures multiples

Rocio Abascal, Michel Beigbeder, Aurélien Bénel, Sylvie Calabretto, Bertrand Chabbat, Pierre-Antoine Champin, Noureddine Chatti, David Jouve, Yannick Prié, Béatrice Rumpler, Eric Thivant

Contact :
LIRIS CNRS FRE-2672
INSA de Lyon
Bâtiment Blaise Pascal
7, avenue Jean Capelle
69621 Villeurbanne Cedex
sylvie.calabretto@liris.cnrs.fr

RÉSUMÉ. *Dans cet article, nous présentons les résultats de recherches collectives[1] sur la «multistructuralité » des documents, menées au sein de l'Institut des Sciences du Document Numérique (ISDN[2]) et financées par la région Rhône-Alpes (Programmes Thématiques 2000-2003). Nos recherches ont abouti à la définition formelle et générique d'un document à structures multiples et d'un certain nombre d'opérations sur ce type de documents (édition, fusion/éclatement, …). Nous illustrons nos propositions sur deux cas d'application : la diffusion des thèses de l'INSA de Lyon et la mise en ligne de la Chronique des Fouilles Archéologiques de l'Ecole française d'Athènes.*

MOTS-CLÉS : *Document numérique, structure documentaire, structures multiples, modélisation de documents, gestion de documents*

Introduction

La problématique de la « multistructuralité » des documents répond à cinq enjeux majeurs dans le domaine de la gestion documentaire :

1. La gestion homogène de différents modèles d'une même information documentaire,

2. La gestion de la cohérence au sein d'un document ou d'une base documentaire,

3. La restitution multiple d'un document,

4. La gestion de l'évolution liée aux différents usages du document (annotations, réutilisation…)

5. La gestion des évolutions structurelles.

D'un point de vue industriel, ce dernier point est particulièrement important. En effet, une grande partie du coût d'un projet documentaire provient de la définition et de la maintenance des structures de documents. Lorsqu'une structure évolue pour un type de document, toutes les applications liées à ce type de document doivent être reprises. De même, les instances de ces documents doivent être souvent transformées, ce qui engendre des coûts très importants, et parfois prohibitifs. La possibilité de gérer plusieurs structures simultanément permettrait de faire évoluer une structure de manière modulaire et souple, et diminuerait les coûts liés à cette maintenance.

Dans le monde documentaire standard, les structures les plus exploitées sont la structure physique et la structure logique. Mais d'autres types de structures liées à la nature et à l'usage des documents sont également représentées et manipulées. Ainsi, de nombreux travaux s'attachent à compléter les structures physiques et logiques par des structures « sémantiques » [NAN 96], [CHA 97], [POU 97] : structure linguistique, discursive [FOU 98], conceptuelle (cf. Web sémantique). Nous pouvons citer en particulier les structures nécessaires à l'adaptation aux utilisateurs dans le cas de restitutions multiples (vue synthétique/développée, vue néophyte/expérimenté, vue français/anglais, combinaison de ces restitutions, etc.). Par ailleurs, la modélisation des hypermédias doit s'intéresser aux structures d'arrangement spatial et/ou spatio-temporel [LEE 97].

Généralement chaque structure est abordée d'une manière individuelle. L'étude globale des multiples structures d'un document et de leurs interactions n'a malheureusement pas fait l'objet d'une grande attention. Dans cet article, nous proposons d'apporter des solutions à la représentation et à la gestion des documents à structures multiples.

La première partie s'attache à proposer un modèle pour les documents multistructurés, et les corpus. Ensuite, nous aborderons les aspects dynamiques du modèle, en particulier, les opérations d'édition, de fusion et

[1] Les laboratoires impliqués sont le LIRIS-INSA de Lyon, LIRIS-Université Lyon1, l'Ecole française d'Athènes, ERSICO – Université Lyon 3, le Centre SIMMO – Ecole des mines de Saint-Etienne et la CNAF-CNEDI de Lyon

[2] http://isdn.enssib.fr/index.htm

[+] *Ecole française d'Athènes*
6 rue Didotou
106 81 Athènes
Grèce

d'éclatement de structures. Ces propositions seront alors illustrées en s'appuyant sur deux corpus : un ensemble de thèses scientifiques et une publication périodique en sciences humaines.

Proposition de modèle

Le modèle proposé a pour double objectif de permettre (*i*) la prise en compte de la multistructuralité des documents, et (*ii*) la modélisation des documents multistructurés et des corpus de façon relativement uniforme. Pour cela, nous définissons tout d'abord un document multistructuré comme un ensemble de structures documentaires mises en correspondance. La notion de multistructure documentaire étend ensuite ce modèle afin de le rendre applicable à des corpus. Nous introduisons enfin la notion de catalogue pour rendre compte de la manifestation d'un corpus sous la forme d'un document.

Document multistructuré

Les structures documentaires sont habituellement modélisées sous la forme d'arbres. De façon plus générale, nous choisissons de nous affranchir des outils de description « technologiques » et de représenter une structure documentaire par un *graphe*, ce qui offre de plus riches possibilités de description.

Définition 1. Une *structure documentaire* est une description d'un document par un ensemble d'éléments en relation les uns avec les autres, au cours ou en vue d'un usage. Mathématiquement : on a un ensemble d'éléments et des relations binaires. Une structure documentaire est donc un multigraphe étiqueté $S = \langle N, L_N, l_N, L_A, A \rangle$ où :

- N est l'ensemble des nœuds du graphe (éléments de la structure)
- L_N est l'ensemble des étiquettes (labels) des nœuds (« contenu » des éléments)[3]
- $l_N : N \rightarrow L_N$ est la fonction qui à chaque nœud associe son étiquette
- L_A est l'ensemble des étiquettes des arcs (« noms » des relations)
- $A \subseteq N \times N \times L_A$ est l'ensemble des arcs (relations nommées entre éléments)
- on ajoute comme contrainte que le graphe doit être connexe[4]

Définition 2. Une *correspondance* entre deux structures est une relation binaire non vide des éléments de la première vers ceux de la seconde.

On note $corr(S_i,S_j)$ l'ensemble des correspondances possibles entre deux structures S_i et S_j. Formellement, $corr(S_i, S_j) = \wp(N_i \times N_j) - \{\varnothing\}$, où $\wp(E)$ désigne l'ensemble des parties d'un ensemble E et $S_i = \langle N_i, L_{Ni}, l_{Ni}, L_{Ai}, A_i \rangle$ et $S_j = \langle N_j, L_{Nj}, l_{Nj}, L_{Aj}, A_j \rangle$.

Définition 3. Un *document multistructuré* est un document structuré dans lequel on considère plusieurs usages possibles et donc plusieurs décompositions structurelles. L'une de ces structures, nommée *structure première*, est constitutive du document en tant qu'unité. Toute autre structure s'appuie sur la structure première par le biais d'une *correspondance* avec celle-ci, directement ou par l'intermédiaire d'une autre structure. Enfin, deux structures quelconques ne peuvent pas être mutuellement en correspondance, ni directement ni indirectement. Formellement, $D = \langle S_0, \Sigma, C \rangle$ où :

- S_0 est la structure première du document,
- $\Sigma = \{S_i \mid i = 1..n\}$ est l'ensemble des autres structures du document,

[3] Les ensembles d'étiquettes L_N et L_A constituent l'ensemble des étiquettes présentes dans le graphe. Ces étiquettes peuvent provenir d'un « vocabulaire » défini extérieurement, bien que ce ne soit pas requis par la définition. On peut faire l'analogie avec la possibilité, en XML, d'écrire un document sans se référer à une DTD.

[4] La contrainte de connexité mérite discussion. Une structure documentaire correspond à la description d'un document en vue d'un usage. Si la structure comportait plusieurs composantes totalement décorrélées, elles constitueraient deux structures distinctes. On peut rapprocher cette contrainte à celle qui, en XML, impose une racine unique à chaque document (et donc la connexité).

- $C = \{C_{ij} \in corr(S_i, S_j)\}$ est l'ensemble des correspondances tel que :
  - $\forall i \neq 0, \exists j < i \mid C_{ij} \in C$ (connexité et convergence),
  - $\forall i \leq j, C_{ij} \notin C$ (absence de circuit).

Remarquons d'abord qu'un document monostructuré est un document multistructuré particulier $D = \langle S_0, \varnothing, \varnothing \rangle$. D'autre part, l'ensemble C des correspondances entre structures définit une relation entre structures qui permet de considérer un *graphe des structures.* Les contraintes imposées à C confèrent à ce graphe certaines propriétés :

- il est connexe et possède comme puits unique S_0 (les correspondances « convergent » toutes vers S_0), puisque toute structure s'appuie directement ou indirectement sur la structure première,
- il ne comporte pas de circuit : en effet, si deux structures sont définies en se faisant mutuellement référence, nous considérons qu'elles constituent une unique structure d'usage.

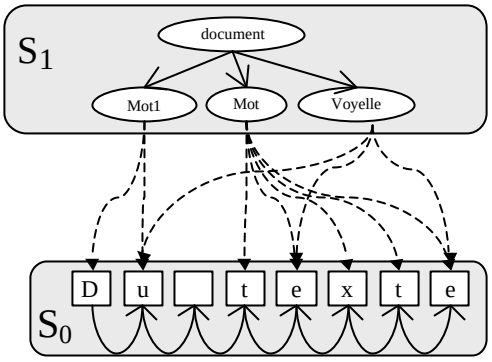

Figure 1 a) Un document bistructuré : la structure S_0 correspond à un texte sans mise en forme, enrichi par la structure S_1.

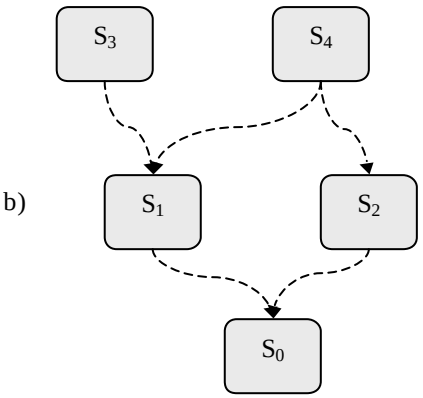

Figure 1 b) Un document multistructuré (seul le graphe des structures est représenté)

Corpus et catalogue

Définition 4. On appelle *corpus* un ensemble de documents multistructurés.

Définition 5. Une *multistructure documentaire M* est un ensemble de structures documentaires mises en correspondance. Formellement, $M = \langle \Sigma, C \rangle$ où :

- $\Sigma = \{S_i \mid i = 1..n\}$ est l'ensemble des structures documentaires
- $C = \{C_{ij} \in corr(S_i, S_j)\}$ tel que $\forall i \leq j, C_{ij} \notin C$ (absence de circuit)

Remarquons que la définition d'un *document multistructuré* (définition 3) vérifie celle d'une *multistructure documentaire* en y ajoutant les contraintes imposant la « convergence » et la connexité du graphe des structures. Par ailleurs, un ensemble de documents multistructurés (corpus) constitue une multistructure documentaire.

Définition 6. Un *catalogue* est une manifestation documentaire d'un corpus. En tant que document, le catalogue possède une multistructure conforme à la définition de document multistructuré. En tant que manifestation d'un corpus, on peut identifier dans sa multistructure les documents constituant ce dernier.

Par exemple, les actes d'une conférence, ou une thèse contenant des documents-images (que l'on peut voir comme une thèse, ou comme manifestation du corpus d'images utilisé par le thésard) sont des catalogues.

Aspects dynamiques

Au niveau de modélisation auquel nous nous situons dans cette partie, extrêmement générique, il n'est bien évidemment possible que de proposer de grandes catégories d'opérations. Celles-ci doivent être précisées pour les différents types de structures et de documents. Leur applicabilité sera en pratique soumise à des contraintes dépendant notamment de la sémantique prêtée aux différentes structures.

Opérations d'édition

L'édition d'un document multistructuré peut donner lieu à l'ajout ou la suppression : (*i*) de nœuds ou d'arcs dans les structures documentaires (la modification de l'étiquette d'un nœud ou d'un arc peut toujours être vue comme un ensemble de suppressions/créations) ; (*ii*) de liens de *correspondance* entre nœuds de différentes structures ; (*iii*) de structures documentaires.

Les opérations d'éditions peuvent modifier les propriétés d'une multistructure documentaire qui lui confèrent le statut de *document* multistructuré (connexité du graphe des structures, convergence des correspondances vers une structure unique S_0). Comme les applications sont en général « orientées document », ceci se traduira généralement par la séparation d'un document en plusieurs, où à l'inverse par la fusion dans un document unique d'éléments provenant de plusieurs autres.

Opérations de fusion/éclatement

On appelle *fusion* la construction d'une structure documentaire à partir de plusieurs autres, et *éclatement* l'opération inverse.

Une structure documentaire étant par définition représentée par un graphe connexe, il n'est pas directement possible d'éclater une structure documentaire valide (i.e. vérifiant les contraintes de connexité), ni d'obtenir une structure documentaire valide par la seule juxtaposition de plusieurs autres. C'est la relation de correspondance entre les nœuds de structures différentes qui permet de résoudre ces problèmes. On envisage en effet que *l'éclatement* de structures sera réalisé en utilisant l'un et/ou l'autre des procédés suivants (cf. figure 3) :

- transformation de certains arcs en relation de correspondance,
- duplication de certains sommets avec une relation de correspondance entre les copies.

Réciproquement, la *fusion* de plusieurs structures en une seule sera réalisée par :

- transformation des certaines relations de correspondance en arcs,
- fusion de certains nœuds en correspondance en un seul.

Il est important de remarquer que ces modifications au niveau des structures sont fondamentalement différentes des opérations d'édition : elles ne consistent pas forcément en un changement effectif du document, mais simplement dans un changement de *point de vue* sur le document. Notamment, c'est ce qui permet d'envisager un catalogue (cf. partie précédente) tantôt comme un document à part entière, tantôt comme la manifestation d'un corpus (sous la forme d'une multistructure documentaire).

Rappelons également que si en théorie tous les éclatements et fusions sont envisageables, dans la pratique l'application à des documents réels restreint les possibilités. Nous présentons dans la suite, deux illustrations instanciant le modèle.

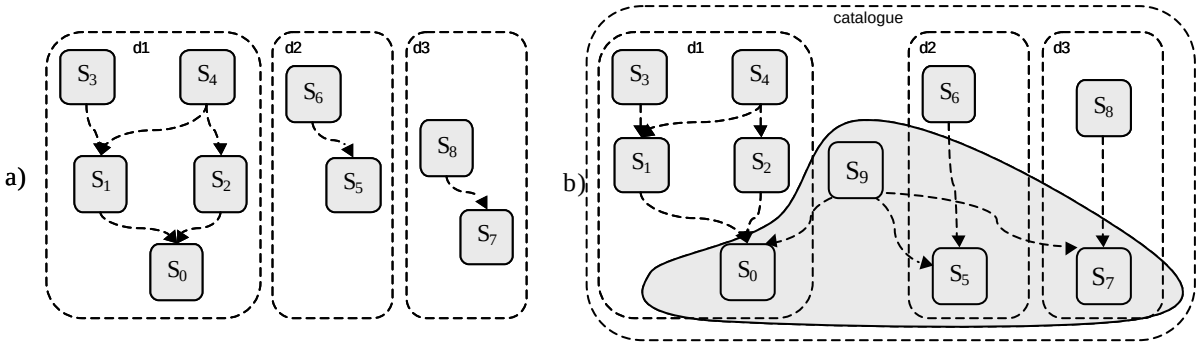

Figure 2 a) Un corpus vu comme une multistructure documentaire **b) Un catalogue** lié à ce corpus. L'ensemble des structures entourées constitue la structure première du catalogue. Les structures S_0, S_5, S_7 et S_9 ne sont représentées que pour indiquer la génèse de cette nouvelle structure première (cf. partie sur les opérations)

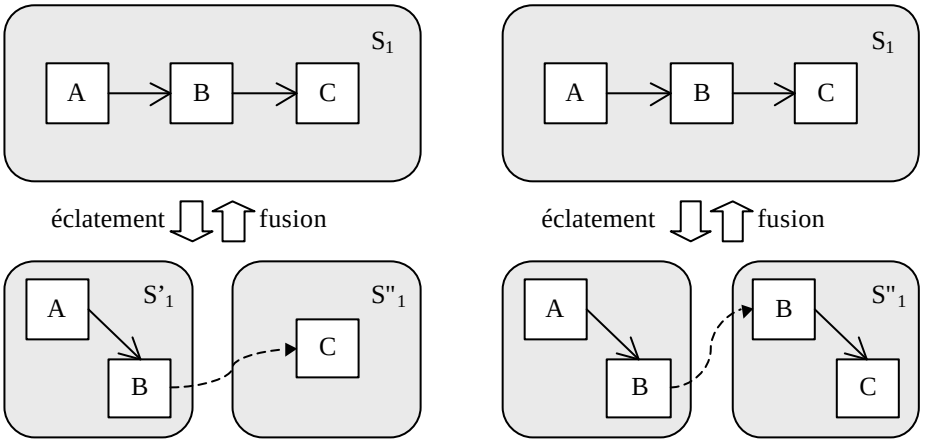

Figure 3 Deux façons d'éclater/fusionner une structure S_1

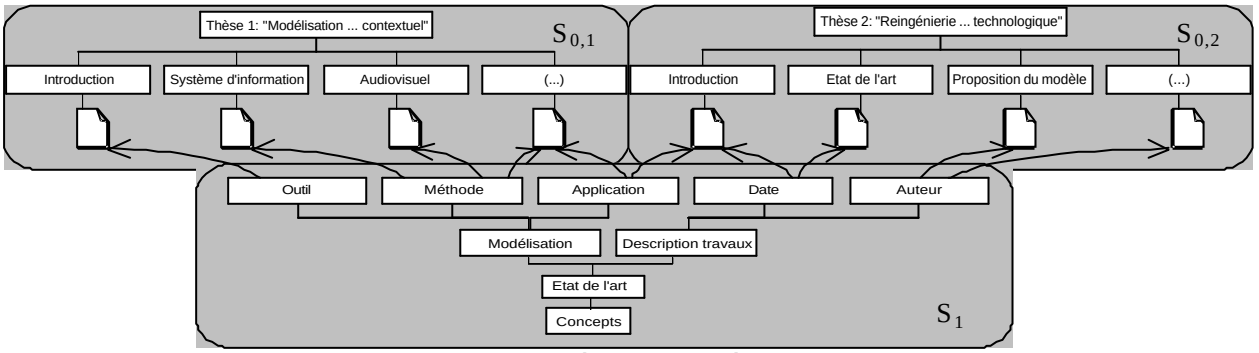

Figure 4 Une multistructure documentaire constituée de deux thèses munies de leur structure logique ($S_{0,i}$) et reliées à une structure « sémantique » commune (S_1).

Illustrations du modèle

Thèses en Sciences de l'Ingénieur

Notre première illustration porte sur un projet de diffusion sous forme électronique des thèses de l'INSA de Lyon. L'objectif de ce projet est d'offrir un accès pertinent à des thèses ou des fragments de thèse.

Dans cet exemple, chaque thèse T_i est caractérisée par sa structure logique $S_{0,i}$ (chapitre, section, texte…). D'autre part, une structure « sémantique » S_1 commune à toutes les thèses du corpus est obtenue par extraction automatique de concepts.

Pour des besoins d'indexation, S_1 est mise en correspondance avec les éléments des structures $S_{0,i}$. Le corpus des thèses constitue donc une multistructure documentaire $M = \langle \Sigma, C \rangle$ où $\Sigma = \bigcup_i \{S_{0,i}\} \cup \{S_1\}$ et $C = \bigcup_i \{C_{1\to 0,i}\}$ (cf. figure 4).

Les documents bistructurés s'appuyant sur une structure logique et la structure « sémantique » correspondent aux thèses indexées. La recherche d'information se ramène alors à la comparaison entre un graphe requête (« domaine », « modèle », « résumé », ...) et les structures des différentes thèses.

Une publication périodique en sciences humaines

Notre seconde étude de cas porte sur la mise en ligne de la *Chronique des fouilles et découvertes archéologiques*[5] (du *Bulletin de Correspondance Hellénique*, revue de l'Ecole française d'Athènes). Le rôle de cette chronique est de signaler chaque année les « nouveautés » archéologiques dans le monde grec.

Structure originelle du corpus

Originellement, la *Chronique* est diffusée sous forme de livraisons annuelles (depuis 1920) constituées (cf. figure 5) :

- d'éléments documentaires (textes courts, photographies, plans) intelligibles indépendamment les uns des autres, mais cependant organisés selon une séquence de lecture $S_{0,i}$ (pour interpréter des abréviations comme «*op. cit.* » ou « *ibid*. »),
- d'une structure bibliographique $S_{1,i}$ permettant de gérer les références internes et externes au corpus, toutes de la forme « BCH n°44, p.409-410 »,
- d'une hiérarchie de sections $S_{2,i}$ donnant une impression de « zoom » géographique. Celle-ci étant globalement invariante d'une livraison à une autre.

Restructuration du corpus

Dès lors que l'on souhaite gérer un réel corpus et non chaque livraison de manière indépendante, il devient nécessaire de restructurer le corpus de la manière suivante (cf. figure 6) :

- concaténer les séquences de lecture dans une séquence globale S_0 (c'est-à-dire en ajoutant une relation de séquence entre le dernier élément de chaque $S_{0,i}$ et le premier élément de $S_{0,i+1}$),
- unir dans le catalogue S_1 du «*Bulletin de Correspondance Hellénique* » la multistructure documentaire des structures bibliographiques $S_{1,i}$,
- fusionner dans une sorte de thésaurus spatial S_2 les hiérarchies de section $S_{2,i}$.

Concernant la fusion, notons qu'il s'agit d'une manipulation de structure particulièrement complexe. Nous devons par exemple tenir compte de changements de nom (Thessalonique/Salonique), de différences de granularité (ville/quartier) ou de changements structurels encore plus radicaux (la Thrace moderne correspond à la Thrace occidentale antique car la Thrace orientale antique est aujourd'hui en Turquie).

Pour manipuler et interroger conjointement ces différentes structures, nous avons été amenés à créer le logiciel libre *Porphyre*[6] (cf. [BEN 02]).

[5] Action concertée incitative « Numérisation des revues en sciences humaines et sociales », Ministère de la Recherche

[6] http://www.porphyry.org (inauguration courant 2003)

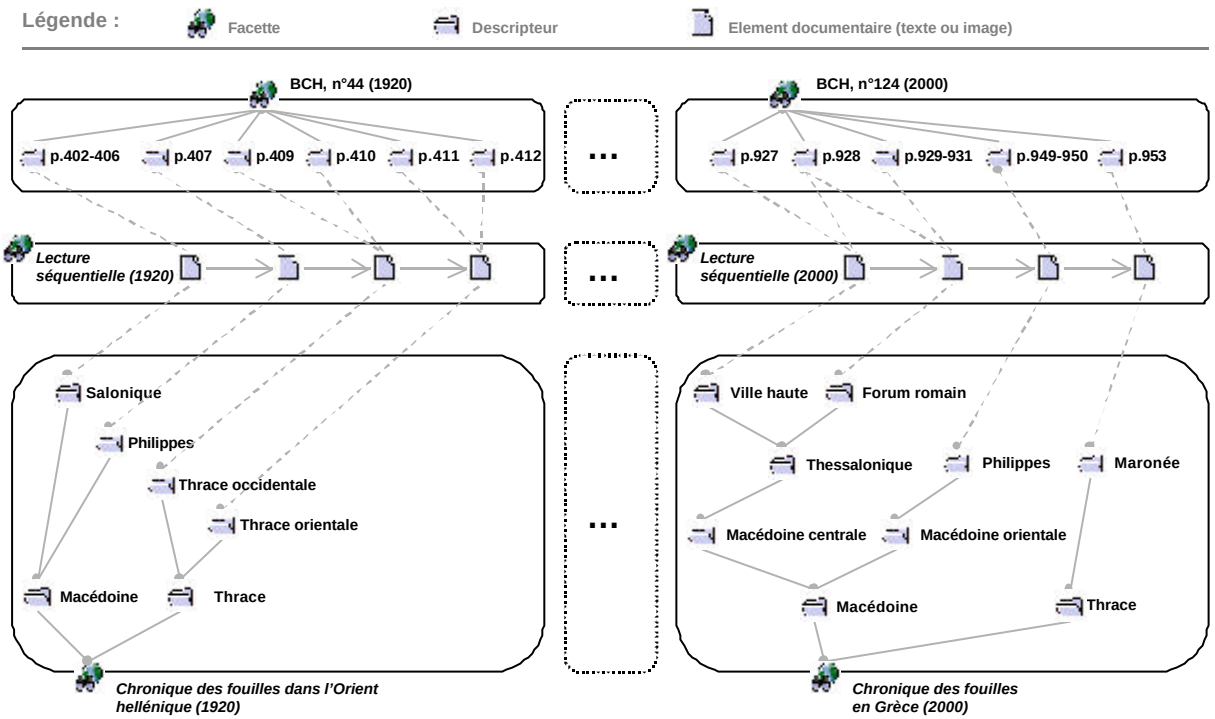

Figure 5 La *Chronique des fouilles* : Extrait de la structure originelle du corpus

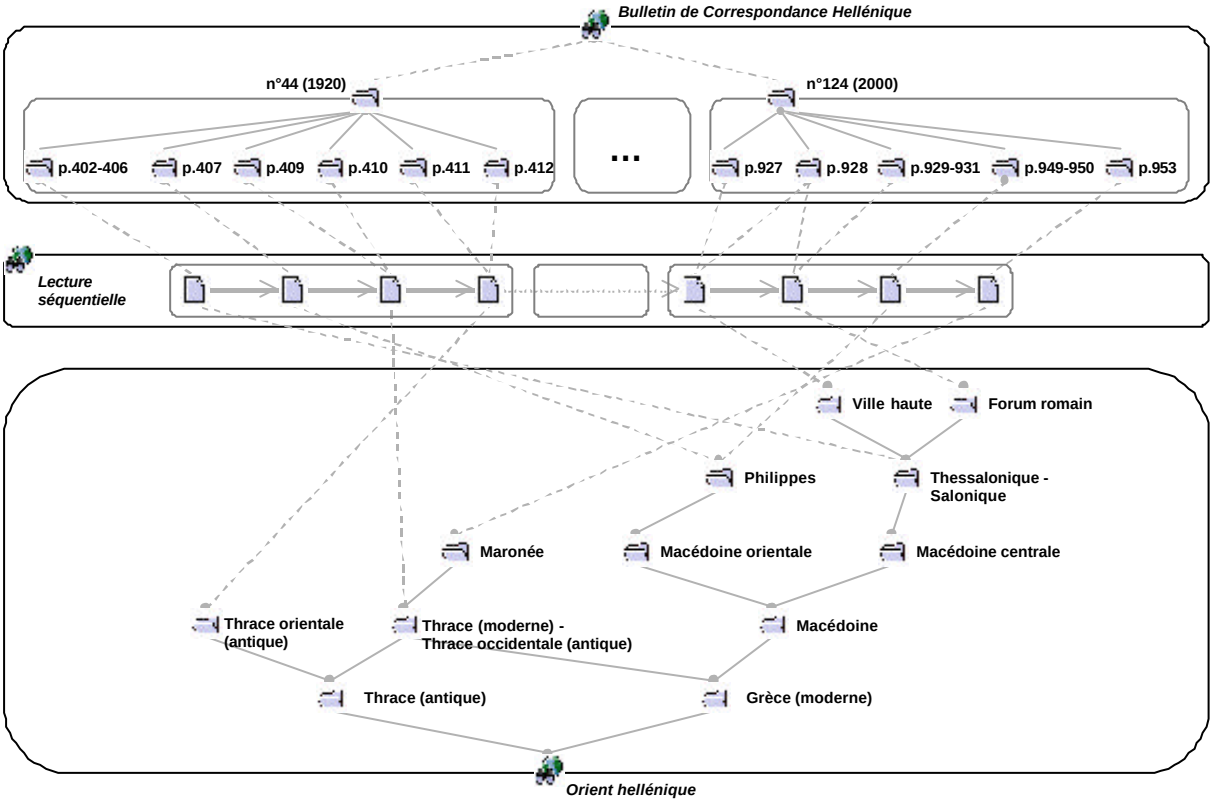

Figure 6 La *Chronique des fouilles* : Extrait de la structure du corpus après restructuration

Travaux apparentés

Des initiatives, des modèles ou des langages abordent déjà cette notion de structures multiples (cf. figure 6). Notre proposition se distingue principalement par :

- une extension de la notion de structure. A l'heure actuelle, de nombreuses applications manipulant des structures de graphes quelconques se voient contraintes d'user de stratagèmes *ad hoc* pour les conformer en arbres (XML ou SGML [ISO 86]) ;
- la possibilité d'établir une relation directe entre les structures sans passer par la structure de base. L'intérêt d'un tel mécanisme paraît évident lorsque l'on souhaite "aligner" les structures d'un ensemble de documents. Par exemple, pour un catalogue donné, traduire une structure DUBLIN CORE en UNIMARC ;
- la définition d'un cadre abstrait de représentation documentaire qu'il s'agit ensuite de spécialiser (instancier) en fonction des médias (texte, image, audiovisuel, etc.) ;
- l'abstraction d'un certain nombre d'usages : restitution, navigation, recherche d'information, annotation, manipulation de documents et de corpus, etc. (contrairement à [NAV 95], [MEC 95] et [FOU 96]).

Bilan et perspectives

La formalisation de la multistructuralité documentaire répond avec pertinence à la question du sens dans les documents numériques, à l'heure de leur interconnexion généralisée par l'utilisation croissante du langage XML et de ses standards associés.

Nous avons proposé dans cet article un modèle formel de documents à structures multiples. Si dans les travaux antérieurs, plusieurs sortes de structures avaient été envisagées, généralement elles n'étaient pas considérées comme parties d'un tout, et étaient utiles essentiellement pour l'indexation et la recherche d'information. L'apport fondamental de notre modèle est de permettre, d'une part l'intégration de tout un ensemble de structures (usuelles, normatives, etc.), d'autre part la représentation et la manipulation de documents multistructurés. De plus, l'approche proposée permet de réfléchir à de nouvelles pratiques documentaires, hypertextes et hypermédias, notamment l'analyse et la gestion (lecture, écriture, annotation, maintenance, réorganisation).

Il est important de noter que la seule contrainte forte concerne l'absence de circuits dans le graphe des structures, cette contrainte est le reflet de la création successive des différentes structures. Par contre aucune hypothèse n'est imposée par notre modèle quant à la nature des structures mises en œuvre. Celui-ci peut donc s'appliquer aux classiques structures physique et logique, mais aussi à toutes les sortes de structures sémantiques, ou encore aux différentes structures temporelles utilisées dans le cadre des documents audiovisuels.

Nos travaux se poursuivent par la mise à l'épreuve du modèle sur d'autres cas d'application. En parallèle, nous affinons la formalisation du modèle, en particulier celle des opérations, avec pour objectif la spécification, dans le cadre de l'ISDN, d'une boite à outils pour la gestion de documents à structures multiples.

Références

[BEN 02] A. BENEL, S. CALABRETTO, A. IACOVELLA, J.-M. PINON. *Porphyry 2001: Semantics for scholarly publications retrieval*. Proceedings of the thirteenth International Symposium on Methodologies for Intelligent Systems [ISMIS]. LNAI #2366. Springer-Verlag, 2002. pp. 351–361.

[CHA 97] B. CHABBAT. *Modélisation Multiparadigme de textes réglementaires*. Thèse de doctorat, LISI. Lyon, décembre 1997, 392 p.

[FOU 96] F. FOUREL. *Modelling Multimedia Structured Documents : A retrieval Oriented Approach*. DEXA Workshop 1996, pp. 179-184

[ISO 86] *Standard Generalized Markup Language (SGML).* International Organization for Standardization (ISO), Information Processing – Text and Office Systems – ISO 8879-1986

[LEE 97] K. LEE, Y.K. LEE, P.B. BERRA. *Management of Multistructured Hypermedia Documents: Its Data Model, Query Language, and Indexing Scheme.* Multimedia Tools and Applications, vol. 4, no. 2, pp. 199–224, Kluwer Academic Publishers, Boston, Massachusetts, March 1997.

[MEC 95] M. MECHKOUR. *A multifacet formal image model for information retrieval.* MIRO final workshop, Glasgow, UK, 1995, pp. 1–12.

[NAN 96] M. NANARD, J. NANARD, ET AL. *La métaphore du généraliste : acquisition et utilisation de la connaissance macroscopique sur une base de documents techniques.* Acquisition et Ingénierie des Connaissances - Tendances actuelles. N. Aussenac-Gilles, P. Laublet, C. Reynaud. Toulouse : CEPADUES, 1996, pp 285–304.

[NAV 95] G. NAVARRO, R. BAEZA-YATES. *A language for queries on structure and contents of textual databases.* ACM SIGIR. – Seattle, USA, July 1995.

[POU 97] L. POULLET, J.M. PINON, S. CALABRETTO. *Semantic Structuring of Documents.* Proceedings of the Third Basque International Workshop on Information Technology , BIWIT'97, Biarritz, July 1997, pp. 118–124.

[TEN 2002] J. TENNISON, W. PIEZ. *The Layered Markup and Annotation Language (LMNL).* In Extreme Markup Languages 2002. August 2002. http://www.extrememarkup.com/extreme/

MODELES	**USAGE**	**STRUCTURE DE BASE**	**AUTRES STRUCTURES**	**RELATIONS ENTRE STRUCTURES**
CONCUR SGML [ISO 86]	Représentation de multiples arborescences SGML.	Séquence de caractères (texte)	Arborescences	Par le texte.
LMNL [TEN 02]	Représentation de structures multiples autorisant les chevauchements d'éléments au sein d'une même structure.	Séquence de caractères (texte)	Graphes d'intervalles	Par le texte.
Modèle de G. Navarro et R. Baeza-Yates [NAV 95]	Indexation textuelle et structurelle de segments de texte.	Séquence de caractères (texte) plus marqueurs de segments	Arborescences (*Vues*)	Par le texte.
EMIR^2 [MEC 95]	Description et recherche d'images selon différents points de vue.	Ensemble de régions disjointes (*Structural view*)	Graphes	Par la structure de base.
Modèle de F. Fourel [FOU 96]	Indexation de documents multimédias.	Arborescence d'éléments logiques (*Structural view*)	Arborescences (*Vues*)	Par la structure de base.
Notre modèle	Gestion (représentation, manipulation ,..) de documents multistructurés	Graphe (*Structure première*)	Graphes	Par la relation de correspondance entre structures quelconques.

Figure 7 *Travaux apparentés*