Modéliser la structuration multiple des documents

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

LIRIS – INSA de Lyon
Bâtiment Blaise Pascal
7, avenue Jean Capelle
696212 Villeurbanne Cedex
Contact : sylvie.calabretto@liris.cnrs.fr

RÉSUMÉ. Dans cet article, nous présentons les résultats de recherches collectives sur la « multistructuralité » des documents, menées au sein de l'Institut des Sciences du Document Numérique (ISDN[1]) et financées par la région Rhône-Alpes. Ces recherches ont abouti à la définition formelle et générique d'un document à structures multiples.
MOTS-CLÉS : Document numérique, structure documentaire, gestion de documents

1. Introduction

La problématique de la « multistructuralité » des documents répond à cinq enjeux majeurs dans le domaine de la gestion documentaire : i/ la gestion homogène de différents modèles d'une même information documentaire ; ii/ la gestion de la cohérence au sein d'un document ou d'une base documentaire ; iii/ la restitution multiple d'un document ; iv/ la gestion de l'évolution liée aux différents usages du document (annotations, réutilisation…) ; v/ 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. Dans le monde documentaire standard, les structures les plus exploitées sont la structure physique et la structure logique. De nombreux travaux s'attachent à les compléter par des structures « sémantiques » (Nanard *et al.*, 1997), (Chabbat, 1997), (Poullet *et al.*, 1997) : structure linguistique, discursive, conceptuelle… 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.

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

2. 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.

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) , $l_N:N\rightarrow L_N$ est la fonction qui à chaque nœud associe son étiquette, L_A est l'ensemble des étiquettes des arcs (« noms » des relations) et $A\subseteq N\times N\times L_A$ est l'ensemble des arcs (relations nommées entre éléments). On ajoute comme contrainte que le graphe doit être connexe.

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 Si et Sj. 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, $\Sigma=\{S_i|i=1..n\}$ est l'ensemble des autres structures du document, et $C=\{C_{ij}\in corr(S_i,S_j)\}$ est l'ensemble des correspondances tel que $\forall i\neq 0,\exists j<i|C_{ij}\in C$ (connexité et convergence) et $\forall i\leq j,C_{ij}\notin C$ (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 : i) il est connexe et possède comme puits unique S_0 (les correspondances « convergent » toutes vers S_0) ; ii) il est acyclique.

À titre d'exemple, on peut considérer un document textuel dont la structure première S_0 est la simple séquence des caractères. Une structure hiérarchique S_1, s'appuyant sur S_0, regroupe ces caractères en mots, paragraphes, chapitres. Deux autres structures S_1 et S_2, s'appuyant également sur S_0, regroupe ces caractères en lignes et en pages, respectivement pour une sortie écran et une sortie papier. Enfin, une structure S_3 propose une séquence des éléments de S_1, différente de celle induite par l'ordre des caractères dans S_0 (par exemple, une chronologie diégétique, différente de la chronologie narrative).

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 $\Sigma=\{S_i|i=1..n\}$ est l'ensemble des structures documentaires et $C=\{C_{ij}\in corr(S_i,S_j)\}$ 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.

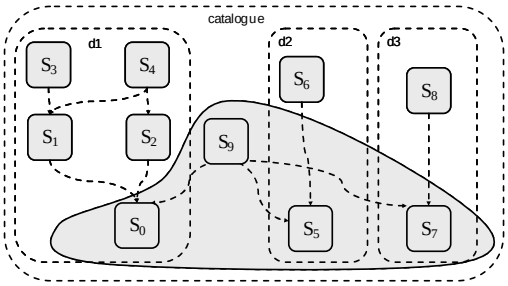

Figure 1. *Exemple de multistructure documentaire*

Dans la figure 1, le corpus constitué des documents d1, d2 et d3 (à l'exclusion, donc de la structure S9), constitue une multistructure documentaire. Si on ajoute une strucure S9 reliant les structures premières des trois documents, on peut considérer l'ensemble S0, S5, S7 et S9 comme une nouvelle structure première d'un catalogue lié à ce corpus, constituant alors en soi un document multistructuré.

3. Conclusion et perspectives

Des initiatives, des modèles ou des langages abordent déjà cette notion de structures multiples : XML ou SGML (SGML, 1986), indexation textuelle et structurelle de segments de texte (Navarro *et al.*, 1995), description et recherche d'images (Mechkour, 1995) et indexation de documents multimédias (Fourel, 1996)... Notre proposition se distingue principalement par : i/ une extension de la notion de structure, ii/ la possibilité d'établir une relation directe entre les structures sans passer par la structure de base, iii/ la définition d'un cadre abstrait de représentation documentaire qu'il s'agit ensuite de spécialiser en fonction des médias (texte, image, audiovisuel, etc.), iv/ l'abstraction d'un certain nombre d'usages (restitution, navigation, recherche d'information, annotation, manipulation de documents et de corpus, etc.). Nos travaux se poursuivent par la mise à l'épreuve du modèle sur plusieurs cas d'application.

4. Bibliographie

Chabbat B. *Modélisation Multiparadigme de textes réglementaires*. Thèse de doctorat, LISI. Lyon, décembre 1997, 392 p.

Fourel F. Modelling Multimedia Structured Documents : A retrieval Oriented Approach. DEXA Workshop 1996, pp. 179-184

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

Mechkour M. A multifacet formal image model for information retrieval. MIRO final workshop, Glasgow, UK, 1995, pp. 1–12.

Nanard M., Nanard J., 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.

Navarro G., Baeza-Yates R. A language for queries on structure and contents of textual databases. ACM SIGIR. – Seattle, USA, July 1995.

Poullet L., Pinon J.M., Calabretto S.. Semantic Structuring of Documents. Proceedings of the Third Basque International Workshop on Information Technology, BIWIT'97, Biarritz, July 1997, pp. 118–124.