Exploitation de la carte dans le document géographique composite

Éric Faurot

Exploitation de la carte dans le document géographique composite

Éric FAUROT

GREYC – UMR 6072
Université de Caen
Département d'informatique, Campus Côte de Nacre
Boulevard du Maréchal Juin
F-14032 Caen Cedex La Rochelle Cedex
eric.faurot@info.unicaen.fr

RÉSUMÉ. *Le document géographique composite est une source importante d'information géographique dont il est nécessaire de considérer l'aspect multi-modal. Cet article présente un modèle sémiologique de la carte géographique. Nous montrons comment celui-ci peut être utilisé en recherche d'information dans le contexte du document composite, à travers un modèle sémantique portant sur l'expression du contraste.*

ABSTRACT. *The geographic composite document is an important source of geographic information for which multimodality should be considered as a primary feature. In this article we present a semiologic model for the map. We show how this model can be usefull for information retreival in composite document, through a semantic model based on contrast.*

MOTS-CLÉS : *Information géographique, document composite, carte géographique*

KEYWORDS: *Geographic information, composite document, geographic map*

1. Introduction

Une grande quantité d'information géographique, et en particulier les aspects sémantiques de cette information par opposition aux données factuelles, se trouve dans des documents produits par les géographes. Alors que les outils de production se développent, atlas et études abondent, et se pose la question de la recherche d'information dans ce type de document et de la valorisation de cette information.

Dans cet article nous présentons très brièvement la problématique de la recherche d'information dans le document géographique composite et les travaux antérieurs. Nous proposons ensuite un modèle sémiologique de la carte géographique, qui permettra d'exploiter ce mode d'expression dans ce contexte. Finalement nous proposons un modèle sémantique basé sur la notion de contraste entre espaces, notion particulièrement importante dans le document géographique, et nous montrons comment la multi-modalité du document peut être mise à profit à travers cette notion.

Cette étude s'inscrit dans le cadre du projet GéoSem[1] dont l'objectif général est de développer des méthodes et des outils, linguistiques et informatiques, d'analyse de contenu de documents géographiques à la fois dans leurs composantes textuelles et imagées (en particulier cartographiques). L'objectif est de permettre à un lecteur-utilisateur de rechercher et de trouver rapidement une information dans un document ou ensemble de documents complexes, et éventuellement de recomposer un nouveau document à partir de plusieurs sources ainsi analysées.

2. Document et information géographique

Le processus de construction de ce type de document peut être décomposé de la manière suivante : Étant donnée une problématique, le géographe émet une ou plusieurs hypothèses. Il détermine ainsi un ensemble de données qu'il doit collecter : autres études, données statistiques... Il procède en suite à l'analyse de ces données, les représente sous différentes formes afin de faire ressortir des faits marquants, et en donne finalement un interprétation. Le document géographique résultant est donc par nature composite. L'auteur utilise différents modes d'expression pour construire et véhiculer l'information : Cartes, graphiques, tableaux pour synthétiser, texte pour expliquer. L'information se trouve ainsi répartie entre ces différentes modalités. Une typologie des différents liens au sein d'un ensemble de documents composites est proposé dans [ALL 96]. La question de la recherche d'information dans ce type de document doit donc à la fois en exploiter la multi-modalité et s'intéresser aux concepts manipulés par le géographe.

Les approches dites *de surface* en recherche d'information ne permettent pas donner une caractérisation fine et précise du contenu sémantique des documents. Ces approches, à la fois sur le texte (à travers les activités de TREC[2]) et en traitement

1. Projet STIC du CNRS
2. http ://trec.nist.gov/

d'images, visent plutôt à la recherche dans de grands ensembles de documents. Les descripteurs de bas niveau utilisés ne prennent pas vraiment en compte l'aspect sémantique [SME 00]. En outre, l'approche véritablement multi-modale est encore timide. La notion de lien entre composantes se limite souvent à l'association simple d'un texte avec une image, sans prise en compte détaillée du contenu. Les différents objets qui composent le document sont abordés du point de vue de leur forme informatique (texte ou image) et non en tant que mode d'expression.

Le travail présenté ici s'inscrit dans la continuité d'une réflexion sur la notion d'information géographique[GAI 01]. Celle-ci est considérée comme un processus articulé autour de trois composantes, l'espace le temps et la thématique. De cette réflexion se dégage la notion d'entité géoréférencée, c'est à dire les espaces évoqués et manipulés par le géographe. Des travaux visant à caractériser et extraire ces entités géoréférencées dans le texte ont déjà été menés [MAL 99, MAT 04]. [MAL 00] propose une étude plus particulière de l'expression des liens d'ordre sémantique dans le document géographique. On s'intéresse ici à la manière de caractériser ces entités géoréférencées dans la carte.

3. Un modèle de la carte

La carte est considérée comme un mode d'expression à part entière au sein du document géographique. Le problème est de définir un modèle de contenu pour ce mode d'expression sur lequel s'appuyer pour proposer des traitements. La délicate question de la sémantique des cartes a été l'objet de plusieurs travaux dans le domaine de la cognition[ROB 85, BAR 97, CAR 00]. Dans [PRA 93], il est montré que la sémantique formelle d'une carte est nécessairement basée sur une structure syntaxique qui n'est pas inhérente à la carte elle-même, mais qui dépend de la tâche de lecture considérée. Dans le cas présent, nous nous intéressons à la carte thématique. Ce type de carte vise à retranscrire la spatialité d'un phénomène quantifié ou qualifié sur un espace. Il est couramment utilisé en géographie humaine. La figure 1 est un exemple de carte thématique.

Nous proposons un modèle qui se dérive en trois aspects, physique, graphique et logique, décrits ci-après. Il s'agit d'un modèle basé à la fois sur des considérations pratiques liées à la forme et aux usages de tels documents, et théorique dans la manière de formaliser le lien entre information et représentation. Ce modèle de contenu se base sur la théorie de sémiologie graphique de Bertin[BER 73].

3.1. *Aspect physique*

L'aspect physique de la carte englobe l'ensemble des informations de bas niveau concernant la forme sous laquelle est donnée cette carte : typiquement le format de fichier, ainsi que d'autres données associées comme le nom du fichier, des spécificités

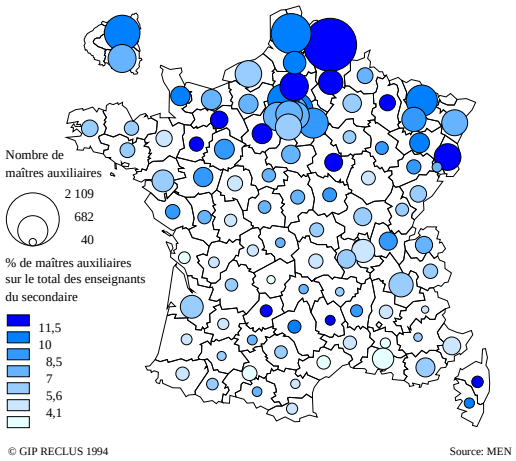


Figure 1. *Exemple de carte thématique*

du format de fichier, etc. Il peut aussi contenir des informations sur le contexte de l'image, par exemple le corpus dont elle est issue.

Cet aspect ne présente pas de formalisme particulier. Il simplement considéré comme un ensemble assez libre d'attributs/valeurs, qui reflète de manière faiblement structuré l'*environnement* de la carte. Cet aspect est en fait principalement dédié à guider et améliorer les traitements ultérieurs sur l'analyse et la mise en valeur du contenu de la carte, en particulier au cours du processus de reconstruction des autres aspects, c'est-à-dire l'expression du contenu dans les autres modèles proposés.

3.2. *Aspect graphique*

Une carte est un objet essentiellement graphique. Lorsque l'on traite des parties textuelles d'un document dans des problématiques de traitement automatique des langues, le texte est presque *naturellement* décomposé de manière atomique en caractères. Les procédés d'analyse peuvent opérer à différents niveaux comme le mot, le groupe, la phrase, ou sur des découpages dans les mots. Quelque soit le ou les niveau(x) pris en compte, il s'agit toujours d'une décomposition en caractères élémentaires. L'aspect graphique de la carte, exprimé dans le modèle proposé ici, reflète la décomposition de cette carte en briques élémentaires, c'est à dire l'*alphabet* gra-

phique, support de l'information. La construction de ce modèle est pilotée par deux facteurs. D'abord par l'observation des usages informatiques, c'est à dire les modèles de représentation des informations graphiques, plus ou moins riches, qui sont habituellement utilisés en visualisation et impression. Le deuxième facteur est l'observation des principes de construction des cartes géographiques, à savoir quels sont les indices graphiques au sens large qui sont pertinents et couramment utilisés dans la construction de ces documents.

Nous considérons une image comme un canevas sur lequel sont empilés des éléments graphiques. Parmi les primitives graphiques retenues nous avons les rectangles, les lignes, les cercles, les polygones, les textes. À chacune de ces primitives sont associés des styles pour le contour (épaisseur, couleur, ...), le remplissage (couleur) et la police de caractère utilisée pour les textes. Il ne s'agit pas ici de définir un modèle graphique vectoriel complet comme SVG ou PDF. Cela dépasserait bien évidemment le cadre de cette étude. L'objectif est en fait de disposer d'un ensemble simple, réduit et cohérent de primitives permettant de décrire au mieux le contenu graphique des cartes.

3.3. *Aspect logique*

L'aspect logique traduit l'organisation de l'ensemble des éléments qui font de l'objet graphique en question une carte. Pour reprendre l'analogie faite avec le texte dans la section précédente, l'aspect logique traduit en fait la *cohérence grammaticale* de la carte. Cet aspect reprend en grande partie les notions décrites dans [BER 73]. Les principaux éléments structurants sont représentés sur la figure 2.

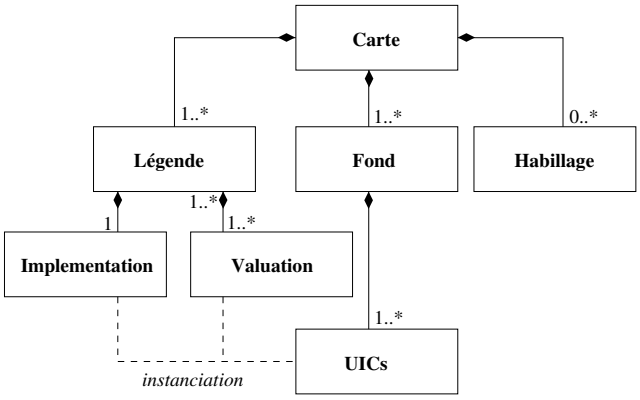


Figure 2. *Aspect logique*

Les éléments d'habillage sont les éléments qui ne constituent pas la représentation de l'information mais qui déterminent le contexte de l'observation. Ici il s'agit du titre,

du sous-titre, du copyright, de la source des données, de l'échelle, etc. Ce sont des éléments à dominante textuelle. Le fond de carte est l'ensemble des éléments graphiques qui précise le contexte géographique de la carte. Il n'apporte pas d'information quant à la mesure elle-même mais délimite l'espace géographique sur lequel porte cette mesure. Si des toponymes sont présents sur ce fond de carte, ils peuvent être considérés comme des éléments d'habillage, dans le sens où ils servent à préciser le contexte de l'observation.

La légende joue un rôle fondamental puisqu'elle spécifie la thématique de la carte et son mode de représentation. Une carte thématique est la construction issue de la mise en relation de plusieurs **composantes**. Une composante est définie de manière générale comme *ce qui varie* dans l'observation, ses dimensions. Le mode cartographique définit implicitement une composante géographique. Les autres composantes thématiques de l'observation sont données par la légende. La définition stricte d'une composante de l'observation est une **valuation**. Elle est décrite par quatre éléments :

– une description de la composante, sous forme textuelle ;

– le type de valeurs que vont prendre les données de cette composante : symbolique, taux, effectif, mesure, pour citer les principaux ;

– un niveau d'organisation, qui peut être qualitatif si les données sont des classes, ordonné si ces classes forme un ordre, ou quantitatif si les données sont des valeurs numériques ;

– une définition, qui spécifie l'ensemble des valeurs possibles ; Il s'agira d'un ensemble de classes ou d'un intervale de valeurs selon le type de la valuation et le niveau d'organisation.

Chaque légende a une et une seule **implémentation**, qui définit la manière dont sont construites graphiquement les associations de valeurs dans l'ensemble des composantes. Au sens de Bertin, il s'agit de définir les variables visuelles qui seront mises en œuvre. C'est la liaison entre les valuations et la représentation graphique des données. Les implémentations auxquelles nous nous intéressons sont les aplats colorés, qui servent principalement à représenter des valuations en classes et les cercles proportionnels qui représentent des mesures ou des effectifs. Ce sont peut-être les deux types d'implémentations les plus couramment utilisés.

Dans le modèle décrit, une carte n'a pas nécessairement une légende unique. En fait, il y a une légende pour chaque association de composantes représentée de manière indissociée sur la carte, c'est à dire quand la représentation des valeurs partage une implémentation. Dans le cas de la figure 1, la légende propose deux valuations, l'une étant un effectif et l'autre un pourcentage organisé en classes ordonnées. L'implémentation se fait en cercles proportionnels colorés.

Les unités d'information cartographique (UICs) sont les représentations graphiques des séries de l'observation. À chaque entité géographique (les valeurs de la composante géographique) est associée une valeur thématique pour chacune des valuations précisées dans la légende. À travers les règles imposées par l'implémentation choi-

sie, cette association donne lieu à une représentation graphique. Sur la carte de la figure 1 les UICs sont les cercles proportionnels sur le fond de carte, chacun symbolisant l'association des valeurs thématiques (par la taille et la couleur couleur) à chaque département. Ces UICs sont l'expression cartographique des entités géoréférencées.

3.4. *Conclusion sur le modèle*

Le modèle proposé reflète bien le fonctionnement intrinsèque à la représentation cartographique décrite dans [BER 73] en spécifiant comment s'organise la carte et où se trouvent les différents indices sémantiques, tout en étant assez souple dans ce qui est représentable. Les cartes peuvent être considérées à la fois en reconstruction (retrouver les aspects graphique et logique à partir de l'aspect physique) ou en génération (générer une carte physique à partir de ces aspects logique et graphique).

4. Modélisation d'une structure sémantique

4.1. *Motivation et modèle*

La notion de contraste est une notion essentielle dans le document géographique. En effet, la géographie humaine peut être définie comme étant l'étude des relations entre espace et société. Elle vise d'une part à décrire et expliquer la manière dont des phénomènes s'inscrivent dans l'espace, et d'autre part, quels sont les espaces remarquables résultant des activités humaines. Le document géographique exprime ainsi de façon récurrente des oppositions et des rapprochements entre espaces (ou entités géographiques) à travers une quantification ou une qualification de ceux-ci. L'extraction de ces structures contrastives est donc importante en terme d'indexation car elles correspondent à une caractérisation sémantique du document. En outre cette notion de contraste n'est pas liée à un mode d'expression particulier, mais se retrouve de manière transverse dans les différentes composantes du document.

D'un point de vue formel, le modèle retenu est assez simple. Il propose de décrire pour une composante donnée (ici un passage de texte ou une carte thématique) un ensemble de relations binaires, notées R, entre deux entités géoréférencées E_1 et E_2. Ces relations sont typées par une valeur symbolique $contraste$ ou $uniforme$. Les entités E_1 et E_2 sont des références à des éléments exprimés dans les modèles associées aux composantes elles-même.

Dans le cas de la composante textuelle, L'extraction de ces structures contrastives s'appuie sur plusieurs principes : l'analyse et langue naturelle des expressions spatiales et leur représentation sémantique[MAL 99, BIL 03a], l'analyse des cadres de discours spatiaux et temporels[BIL 03b], et enfin l'analyse de l'organisation du texte en terme de structures réhtoriques. Ces points sont abordés plus en détails dans [WID 04].

4.2. *Cas de la carte*

La caractérisation des contrastes sur la carte se base sur le modèle sémiologique décrit dans la section précédente. Ici, les contrastes vont porter sur les UICs, qui sont des entités géoréférencées. Il s'agit d'identifier sur la carte des groupes d'UICs qui sont homogènes. La figure 3 montre des contrastes entre des groupes d'UICs qui ont été définis simplement en ségmentant les classes proposées dans la légende en deux groupes, qui constituent deux nouvelles classes. Les éléments connexes et dont la valeur thématique associée tombe dans la même classe sont regroupés. Dans le formalisme retenu, les groupes de la même classe sont en relation d'uniformité deux à deux, et en relation contrastive avec les autres.

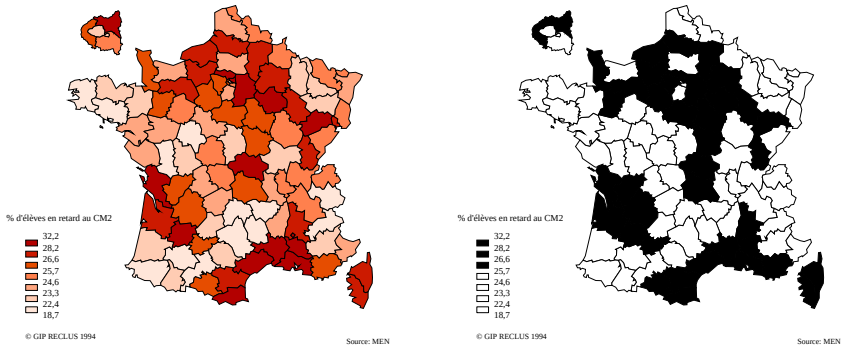
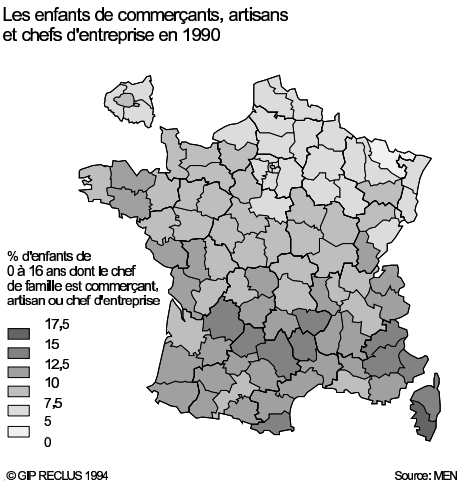


Figure 3. *Contraste sur la carte*

Il s'agit ici d'une possibilité parmi d'autres. En fait il y a différentes manières d'envisager la constitution de ces groupes. La connexité peut être défini sur les éléments graphiques ou sur leurs pendants géographiques (ici les départements) s'ils sont identifiés. La distance entre les éléments peut être prise en compte. Elle doit d'ailleurs l'être dans le cas des cercles proportionnels. Le choix du découpage en classes *utiles* de la légende peut être basé uniquement sur la légende elle-même, comme c'est la cas ici, en fixant un seuil prenant éventuellement en compte la progression du dégradé de couleur. Ce découpage en classe peut aussi être guidé par les UICs, de manière à imposer une certaine répartition du nombre d'éléments dans chaque classe. Finalement, l'appartenance d'une UIC à une classe peut être déterminée non pas uniquement par sa valeur thématique, mais en prenant aussi en considération les UICs voisines. Ainsi, un élément isolé peut être absorbé par l'ensemble de ses voisins formant un groupe plus pertinent. Enfin la recherche des groupes peut être basée sur des principes purement perceptifs en se basant exclusivement sur l'aspect graphique du modèle.



La répartition des enfants de commerçants, artisans et chefs d'entreprise partage clairement la France en deux : la moitié méridionale, où ils sont relativement nombreux, avec 10 à 15% des enfants, et la France du Nord où leur proportion va de 5 à 10%, Bretagne exceptée.

Figure 4. *Une carte et le texte associé*

4.3. *Perspective d'une approche couplée*

L'exemple suivant montre comment l'interprétation conjointe des analyses données par les deux composantes texte et carte peuvent permettre d'enrichir l'information extraite. En effet, le point de vue de l'auteur est véhiculé par l'ensemble des composantes utilisées de manière complémentaire.

La figure 4 présente une carte et un texte qui lui est associé dans le document. On suppose ici que cette association est avérée. La figure 5 montre les structures contrastives résultant de l'analyse des deux composantes. Sur la carte quatre zones semblent pertinentes : Le Nord-Est $M1$ où le phénomène est relativement faible, la Bretagne $M4$ et la partie Sud $M3$ où le phénomène est fort et la partie centrale $M2$. Dans le texte, l'auteur organise le phénomène selon deux zones Nord/Sud $T1$ et $T2$ dont la limite est floue. On suppose que l'analyse textuelle n'est pas en mesure de fournir l'exception de la Bretagne. Le système envisagé permettrait d'obtenir l'interprétation schématisée par le troisième hexagone. La relation de contraste déterminée sur la carte entre $M2$ et $M1$ est révisée en considérant qu'elles constituent en fait le groupe $T1$ du texte ($C1$). La zone sud $T2$ est identifiée à $M3$ ($C2$), précisant ainsi la limite floue du texte. L'exception de la Bretagne est conservée car elle est en relation d'uniformité avec $M3$.

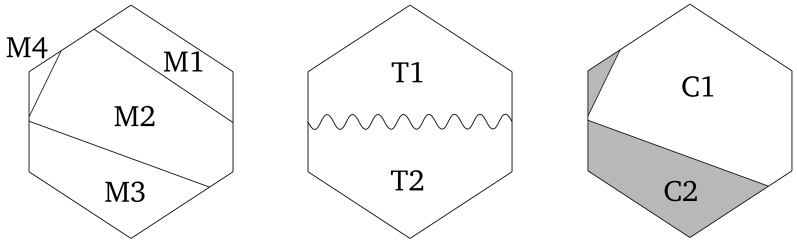


Figure 5. *Contrastes suggérés par la carte, le texte et le croisement des deux*

La mise en œuvre d'un tel système requiert l'utilisation d'un moteur de raisonnement spatial, couplé avec un SIG. Les formalismes nécessaires à l'implémentation doivent encore être spécifiés. Les perspectives sont néanmoins intéressantes. Ce système permettrait notamment d'affiner les analyses en précisant le texte souvent sous-spécifié grâce au caractère exhaustif de la carte qui, en contre-partie, ne propose pas de *grille de lecture* pour son interprétation.

5. Conclusion

Le document géographique composite est une source précieuse d'information, mais en tant que source de données non structurées, son exploitation requiert une réflexion sur la notion d'information géographique et son expression dans ce type de document. On souhaite en particulier mettre à profit la multi-modalité.

Plusieurs travaux ont permis de mieux caractériser la notion d'information géographique et son expression dans la composante textuelle. En vue de pouvoir procéder à une exploitation fines des cartes, nous avons présenté un modèle de contenu pour la carte géographique thématique. Nous avons en outre défini une structure sémantique du document géographique autour de la notion de contraste, qui se retrouve dans l'ensemble des composantes du document. Nous avons montré comment la multi-modalité peut être exploité par le biais de cette structure.

Les travaux actuellement en cours consistent en une consolidation des modèles et des processus d'analyses, et visent à une plus forte intégration des résultats sur les différents modes d'expression. Une collaboration avec des géographes va nous permettre de valider les principes développés en proposant des outils d'aide à la recherche d'information dans un corpus géographique.

6. Bibliographie

[ALL 96] ALLAN J., « Automatic Hypertext Link Typing », *Hypertext 96, Washington, D.C., ACM*, March 1996, p. 42–52.

[BAR 97] BARKOWSKY T., FRESKA C., « Cognitive Requirements on Making and Interpreting Maps », *Spatial information theory : A theorical basis for GIS*, p. 347–361, S.Hirtle & A.Frank, 1997.

[BER 73] BERTIN J., *Sémiologie Graphique*, Mouton & Cie., 2nd édition, 1973.

[BIL 03a] BILHAUT F., CHARNOIS T., ENJALBERT P., MATHET Y., « Passage extraction in geographical documents », *Proceedings of New Trends in Intelligent Information Processing and Web Mining*, Zakopane, Poland, 2003.

[BIL 03b] BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., DRAOULEC A. L., MATHET Y., MIGUET H., PÉRY-WOODLEY M.-P., SARDA L., « Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique », *Proceedings of Traitement Automatique du Langage Naturel (TALN)*, Batz-sur-Mer, France, 2003.

[CAR 00] CARRIÈRE J., « Le paradigme de la communication cartographique : la hierarchisation du "moi" et du "ici" », XYZ, Ed., *Les ancrages du corps propre*, p. 87–110, Nycole Paquin, 2000.

[GAI 01] GAIO M., « *Traitements de l'Information Géographique : Représentations et Structures*, Dossier d'Habilitation à Diriger des Recherches, Université de Caen », 2001.

[MAL 99] MALANDAIN N., GAIO M., « Extraction d'Unités d'Information Géographiques dans de Documents Composites », PRODUCTIONS E., Ed., *Document Electronique Dynamique, CIDE 99*, 1999.

[MAL 00] MALANDAIN N., « La Relation Texte/Image, Essai de Modélisation dans un Corpus Géographique », Thèse de doctorat, Université de Caen, 2000.

[MAT 04] MATHET Y., CHARNOIS T., ENJALBERT P., BILHAUT F., « Geographic reference analysis for geographic document querying », *In Proceedings of Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT)*, Edmonton, Alberta, Canada, 2004.

[PRA 93] PRATT I., « Map Semantics », FRANK A., CAMPARI I., Eds., *Spatial Information Theory : A Theoretical Basis for GIS*, vol. 716, Springer-Verlag, Berlin, 1993.

[ROB 85] ROBINSON A. H., *The look of maps : an examination of cartographic design*, The University of Wisconsin, 2nd édition, 1985.

[SME 00] SMEULDERS A. W. M., WORRING M., SANTINI S., GUPTA A., JAIN R., « Content-Based Image Retrieval at the End of the Early Years », december 2000.

[WID 04] WIDLÖCHER A., FAUROT E., BILHAUT F., « Multimodal Indexation of Contrastive Structures in Geographical Documents », *In Proceedings of RIAO : Coupling approaches, coupling media and coupling languages for information retrieval*, Université d'Avignon, 2004, p. 555–570.