BAKILLAH MOHAMED

DÉVELOPPEMENT D'UNE APPROCHE GÉOSÉMANTIQUE INTÉGRÉE POUR AJUSTER LES RÉSULTATS DES REQUÊTES SPATIOTEMPORELLES DANS LES BASES DE DONNÉES GÉOSPATIALES MULTIDIMENSIONNELLES ÉVOLUTIVES

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en sciences géomatiques
pour l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DE FORESTERIE ET DE GÉOMATIQUE

UNIVERSITÉ LAVAL
QUÉBEC

2007

Résumé

Dans le domaine forestier, la gestion des ressources naturelles se base sur les données recueillies lors des inventaires portant sur la représentation spatiale d'un même territoire à différentes époques. Au fil des inventaires, l'évolution naturelle, les interventions humaines, l'évolution des modes d'acquisition, des spécifications et des normes forestières créent une hétérogénéité spatiale et sémantique entre les différentes bases de données. Dans un processus décisionnel, ces données et spécifications sont structurées d'une façon multidimensionnelle dans des cubes de données géospatiales. Par conséquent, la structure multidimensionnelle est également amenée à évoluer, ce qui affecte la réponse aux requêtes spatiotemporelles. Dans le domaine forestier, la problématique de l'évolution de structure se traduit par l'impossibilité d'effectuer des analyses spatiotemporelles, par exemple sur l'évolution du volume de bois de certaines essences ou l'évolution des épidémies, affectant directement la prise de décision dans la gestion forestière. Cette problématique exige de concevoir de nouvelles solutions capables de préserver les liens entre les membres des différentes structures. Cependant, les solutions proposées ne tiennent pas compte de manière explicite et simultanée de l'évolution sémantique et géométrique de la structure. Afin d'apporter une solution plus adaptée aux réalités des phénomènes spatiotemporels, nous avons développé une approche géosémantique intégrée pour la gestion de l'évolution de la structure du cube afin d'ajuster la qualité de la réponse à la requête spatiotemporelle et ainsi offrir un meilleur support à la prise de décision. L'approche proposée définit une méthode de rétablissement des liens entre des versions du cube. Sur le plan sémantique, nous rétablissons les liens en employant une fonction de similarité sémantique basée sur l'ontologie et qui tient compte du plus fin niveau de définition des concepts. Au niveau géométrique, notre approche se base sur une méthode d'indexation QuadTree pour constituer une matrice de correspondances spatiales entre les géométries des différentes époques. Les liens résultants sont intégrés dans une méthode de transformation matricielle afin de pouvoir répondre d'une manière plus adaptée à des requêtes spatiotemporelles.

Remerciements

Ce travail de mémoire a été effectué au sein du laboratoire CRG (Centre de Recherche de Géomatique de l'Université Laval, Québec). Que soit ici remercié la direction du département, en particulier l'ex-directeur du département Monsieur Jean-Jacques Chevalier et Madame Carmen Couture responsable du secrétariat aux études supérieurs, ainsi que Madame Jacynthe Pouliot, l'actuelle directrice des études supérieurs du département. Mes remerciements s'adressent également à la Chaire Industrielle de Bases de Données Géospatiales du Professeur Yvan Bédard pour le financement de cette recherche. La direction de ce mémoire a été assurée par le Professeur Mir Abolfazl Mostafavi dont le rôle a été déterminant pendant ces deux années de maîtrise. Il a su me guider par de judicieux conseils et il a su être patient et confiant durant les moments les plus difficiles. La co-direction a été menée par le Professeur Yvan Bédard à qui je suis très reconnaissant pour le temps qu'il m'a consacré, la chance qu'il m'a accordé pour faire partie de son groupe, et les judicieux conseils pour le suivi de mes travaux.

Je remercie chaleureusement le Docteur Jean Brodeur d'avoir été l'examinateur de ce mémoire.

Je tiens également à adresser mes remerciements à tous les professionnels de la Chaire Industrielle de Bases de Données Géospatiales du professeur Yvan Bédard qui m'ont épaulé tout au long de cette recherche. J'adresse aussi mes remerciements à tous ceux et celles que j'ai côtoyé pendant ces deux années. Je remercie grandement toute ma famille et mes amis à qui je dois tout et qui m'ont aidé de près ou de loin à la réalisation de ce mémoire.

Un grand merci à toi Jessica et que Dieu te protège.

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1 : Introduction

1. 1 Mise en Contexte

Le processus de prise de décision exige de pouvoir faire des prévisions sur le comportement des phénomènes soumis à l'analyse ainsi que de pouvoir analyser les données sur une période de temps significative. Par exemple, dans le domaine forestier, les outils d'analyse doivent permettre de questionner la base de données sur le comportement, entre différentes époques, de mesures telles que le volume de bois, par rapport à certaines essences, peuplements forestiers, densité de couvert, etc. Les entrepôts de données apparaissent spécialement désignés pour offrir une aide à la décision puisqu'ils contiennent non seulement des données actuelles mais également des données provenant de plusieurs époques, et sont conçus pour permettre le processus d'analyse de ces données. Les outils OLAP « sont une catégorie de logiciel spécialement conçus pour l'exploration rapide et facile des données multidimensionnelles composées de plusieurs niveaux d'agrégation » (Caron, 1998). La structure multidimensionnelle qui les caractérise est en opposition avec l'approche transactionnelle car c'est une représentation plus près de la réalité envisagée par l'utilisateur des données (Codd et al. 1993). De plus, cette dernière facilite la réponse aux requêtes (Kimball, 2002a).

De nombreux travaux proposent des modèles multidimensionnels pour les bases relationnelles (Agra, 1997; Gyss, 1997; Lehn, 1998). Ces travaux exposent différentes applications pour des applications de gestion classiques (transactionnelles) mais ne permettent pas de répondre complètement aux exigences des applications actuelles, telles que les applications médicales (Pederson, 1999). En effet, ces dernières nécessitent des modèles plus riches que les modèles basés sur l'approche relationnelle, afin de gérer des données complexes (spatiales). Ce qui nous ramène à l'approche multidimensionnelle qui introduit de nouveaux concepts comme les dimensions, les mesures et les faits.

Dans un cube, les dimensions sont les thèmes d'analyse thématiques du modèle (Rivest et al, 2005) ; elles regroupent des données descriptives, spatiales ou temporelles (Bédard et al, 2001), par exemple les *catégories de produits*, la *géographie*, et le *temps*. Les dimensions

sont définies par le schéma de la dimension, qui décrit les niveaux de la hiérarchie de la dimension, ainsi que par le schéma des instances contenant les membres de la dimension. Les faits sont les objets de l'analyse (ex : nombre de professeur, surface, total des ventes, etc.).

Dans les modèles multidimensionnels, ces faits sont considérés comme dynamiques, et les dimensions sont statiques, alors qu'en réalité les dimensions sont également sujettes à changer (Cabbibo et al. 1998; Kimball, 1996; Lehner, 1998). La structure multidimensionnelle peut subir des évolutions au niveau du schéma des dimensions et du schéma des instances ; cette problématique est assez bien documentée (Kimball, 1996). Les membres de la structure décrits par des attributs descriptifs et des données spatiales peuvent également subir des évolutions sur le plan sémantique et géométrique.

L'évolution de la structure est causée par des changements dans les méthodes de classification, des changements du format et de la sémantique des données, l'évolution géométrique des objets, etc. En effet, les entrepôts de données intègrent des données qui proviennent de sources hétérogènes, et puisque ces sources sont autonomes, elles peuvent évoluer dans le temps, indépendamment l'une de l'autre, et indépendamment de l'entrepôt de données qui les intègre. La principale conséquence de l'évolution de la structure des dimensions est que les faits d'analyse peuvent être représentés dans différentes structures résultantes de cette évolution. Par conséquent, des requêtes exécutées à différents points dans le temps peuvent fournir des résultats différents et parfois contradictoires (Thurnheer, 2000). Le fait qu'un objet possède plusieurs représentations structurelles empêche également d'agréger les données et de les comparer dans le temps.

Pour tenir compte de l'évolution de la structure multidimensionnelle, il est nécessaire de propager les changements de ces dernières à l'entrepôt de données. Cependant, les entrepôts de données et leurs technologies associées (OLAP), bien qu'ayant pour fonction de gérer des données évolutives, ne sont pas aptes à gérer les évolutions structurelles que sont les changements dans les dimensions et les hiérarchies, et les changements dans les schémas. Cette incapacité est due au fait que, dans le modèle multidimensionnel, toutes les dimensions sont considérées comme orthogonales entre elles, ce qui implique que toutes les dimensions sont indépendantes du temps.

La problématique de l'évolution de la structure multidimensionnelle a été l'objet de plusieurs travaux dans le domaine des entrepôts de données. Les principales solutions apportées sont celles des mises à jour (Quix, 1999; Blaschka, 2000) et du versioning (Balmin et al., 2000 ; Mendelzon, Vaisman, 2002 ; Eder et Koncilia, 2002 ; Body et al. 2002 ; Morzy, Wrembel, 2004). Les approches de mise à jour consistent simplement à ne conserver que la structure la plus récente pour représenter les données ; elles ne sont pas suffisantes pour répondre aux besoins d'analyses temporelles, bien qu'elles permettent la comparaison des données recueillies à différentes époques dans différentes structures. Les approches de versioning ont été conçues dans le but de conserver l'historique de la structure sous forme d'un ensemble de versions ; certaines proposent des méthodes de mapping entre les versions, lesquelles permettent de retracer l'évolution des membres de la structure multidimensionnelle. Cependant, nous constatons que parmi ces approches, l'évolution sémantique et géométrique n'est pas considérée.

Cependant, la problématique de l'évolution n'est pas seulement propre au modèle multidimensionnel ; elle se manifeste à tous les niveaux de représentation (conceptuel, physique, logique). Elle est également le sujet de plusieurs travaux dans le domaine des ontologies (Kalfoglou, Shorlemmer, 2002) où l'approche du versioning y a également été adaptée. Les solutions qui nous ont intéressées se basent sur la sémantique des concepts et emploient une mesure de similarité pour identifier les relations de mapping entre les ontologies.

1. 2 Problématique

Dans le contexte de notre recherche, nous disposons des données qui proviennent de quatre inventaires forestiers, chacun de ceux-ci étant représenté par un cube de données multidimensionnelles. La modification des modes de classification et de la définition (sémantique) des membres du schéma des instances des dimensions, ainsi que le redécoupage spatial des entités géographiques composant les dimensions spatiales ont pour conséquence que les structures des ces cubes sont différentes. Chaque cube représente donc une version de la structure à une époque donnée.

- Suite à l'évolution de la structure multidimensionnelle, les liens sémantiques et géométriques ne sont pas conservés entre les différentes versions du cube de données géospatiales.

- L'absence de liens entre les membres des cubes de données géospatiales qui ont évolués fait en sorte que les résultats des requêtes temporelles (portant sur plusieurs versions du cube) peuvent être faussés. Par exemple, si la définition attribuée à l'essence *bouleau blanc* a évolué entre deux époques, les résultats obtenus concernant le volume de bois de l'essence *bouleau blanc* selon chaque époque ne représentent pas la même réalité et, par conséquent, ne sont pas comparables.

- L'absence de liens entre les membres des cubes de données géospatiales qui ont évolués fait en sorte que les résultats des requêtes temporelles (portant sur plusieurs versions du cube) peuvent être impossible à obtenir. Par exemple, les membres des dimensions spatiales des cubes de données forestières sont des zones du découpage forestier dont la géométrie a évolué et qui n'existent que dans une époque donnée. Il est, par exemple, impossible de connaître le comportement du volume de bois du peuplement 1868 défini en 1984 entre 1984 et 1992 puisque ce peuplement n'existe qu'en 1984.

En conséquence, l'évolution de la structure multidimensionnelle et l'absence de liens (sémantiques et géométriques) entre les membres des différentes structures qui s'ensuit affecte directement la réponse à la requête et nuit à la qualité de l'information et au processus d'aide à la décision.

1. 3 Objectifs

1.3.1 Objectif général

L'objectif général de ce mémoire de maîtrise est d'ajuster la réponse aux requêtes spatio-temporelles dans les cubes de données géospatiales en prenant en considération l'évolution de la structure sémantique et géométrique dans les bases de données multidimensionnelles.

1.3.2 Objectifs spécifiques

1. Évaluer et proposer une approche pour le rétablissement des liens sémantiques entre les cubes de données géospatiales.

2. Évaluer et proposer une approche pour le rétablissement des liens géométriques entre les cubes de données géospatiales.

3. Développement d'une approche géosémantique pour le traitement des requêtes spatio-temporelles dans les cubes de données géospatiales et application puis évaluation de l'approche proposée dans le contexte forestier.

1. 4 Méthodologie

Le problème de l'évolution de la structure multidimensionnelle a été soulevé en 1996 par Kimball qui a établit les fondements des solutions qui ont donné naissance à deux grandes approches : les mises à jour, qui consistent à modifier la structure afin de la rendre actuelle ainsi que les bases de données à versions, dont l'objectif principal est de conserver l'historique de la structure. Toutefois, ces deux approches ne règlent que partiellement le problème de l'évolution de la structure. Seules quelques solutions (Eder et Koncilia, 2002 ; Body et al. 2002 ; Golfarelli et al. 2004 ; Morzy et Wrembel, 2004) permettent d'établir des liens entre les versions de la structure, ce qui est essentiel pour pouvoir répondre correctement à des requêtes qui portent sur plusieurs versions. Cependant, aucune de ces solutions ne tient compte de l'évolution sémantique et géométrique de la structure multidimensionnelle. Pour résoudre ce problème, nous proposons une nouvelle approche intégrée géosémantique qui prend en compte à la fois les problèmes sémantiques et géométriques engendrés par l'évolution de la structure multidimensionnelle. La méthodologie suivie dans ce projet de recherche se divise en cinq phases (voir figure 1.2):

La première phase est la recherche préliminaire qui regroupe la définition du sujet de recherche, la problématique ainsi que la recherche bibliographique. La recherche bibliographique vise à cerner, dans un premier temps, un ensemble de sujets liés aux problèmes d'évolution de la structure multidimensionnelle, plus précisément, elle porte sur les thèmes suivants : les concepts fondamentaux des cubes de données géospatiales, la

typonymie des évolutions de la structure multidimensionnelle, le concept de temps valide, les méthodes de mises à jour et de versioning des bases de données géospatiales multidimensionnelles et les approches de mapping entre les cubes. Une seconde partie de la recherche bibliographique porte sur les thèmes qui relèvent du domaine des mapping entre les ontologies et des modèles de similarité sémantique, ces derniers permettant d'établir les liens entre deux concepts dépendamment de la nature des deux concepts (Rips et al., 1973 ; Tversky, 1977; Cardelli, 1984 ; Miller et Charles, 1991; Lee et al., 1993; Brodeur et al., 2004; Mostafavi, 2006). La dernière partie de la recherche évalue les solutions pour le rétablissement de liens géométriques, en particulier les méthodes d'indexation de données spatiales qui permettent de former une représentation commune pour différentes géométries.

La seconde phase est celle de l'étude des données, issues des inventaires de la forêt de Montmorency, de l'évolution de la structure des cubes de données forestières et des exigences posées par ce contexte. Elle comprend l'analyse préliminaire des données qui seront utilisées pour évaluer notre approche. Nous y établissons ensuite les parallèles entre le cube de données multidimensionnel et l'ontologie du cube sur les plans de la représentation et de l'évolution.

La troisième phase traite du rétablissement des liens sémantiques et géométriques entre les membres des schémas d'instances des différentes versions du cube. Pour le rétablissement des liens sémantique, nous avons évalué, en fonction de notre contexte, la pertinence du modèle *Matching Distance* (Rodriguez, 2000) afin de pouvoir l'adapter aux exigences du contexte choisi, qui comporte des données de types complexes (textes et intervalles), et à le rendre plus flexible et précis ; la mesure de similarité redéfinie au niveau théorique est appliquée non seulement au niveau des concepts, comme dans le cas du modèle Matching Distance, mais aussi au niveau des propriétés des concepts afin d'affiner la mesure, ce qui est nécessaire pour obtenir plus de précision. Le modèle ainsi redéfini, appliqué pour comparer les instances, permet de formaliser une fonction de mapping qui permet de constituer des matrices de correspondances sémantiques qui mettent en relation les membres des schémas des instances de dimensions de deux différents cube de données géospatiales. Pour le rétablissement des liens géométriques, les géométries des membres

des dimensions spatiales des différents cubes sont représentées par des entités spatiales communes au moyen d'une méthode d'indexation *QuadTree* appliquée à partir du module Oracle Spatial. Une fois indexées, les surfaces des membres des dimensions spatiales peuvent être comparées pour constituer des matrices de correspondances géométriques entre les dimensions spatiales des cubes.

La quatrième phase consiste à développer une approche géosémantique intégrée permettant de traiter les requêtes spatio-temporelles en fusionnant les approches de rétablissement de liens sémantique et géométriques avec une méthode de transformation matricielle du cube. La méthode de transformation matricielle permet de transformer les mesures d'une version à l'autre afin de répondre à des requêtes évolutives.

La dernière phase consiste à développer une application afin de tester notre approche avec les données de la forêt de Montmorency ; nous étudierons, dans un premier temps, le comportement du modèle de similarité sémantique et ses performances par rapport au modèle original Matching Distance, puis la pertinence de deux approches probabiliste de prise en compte du contexte dans le modèle de similarité. Puis les résultats des requêtes seront aussi présentés afin de montrer la validité et l'applicabilité de notre approche et éventuellement l'améliorer pour aider au processus de prise de décision.

1. 5 Contenu du mémoire

Ce mémoire se compose de cinq chapitres, excluant l'introduction : le chapitre 2 décrit les concepts de base du modèle multidimensionnel ainsi que la problématique de l'évolution de la structure multidimensionnelle et les impacts de ces évolutions sur les résultats des requêtes spatio-temporelles. Les différents types d'approches proposés pour gérer l'évolution de la structure sont discutés en fonction de leur pertinence à résoudre les différents types d'évolution de la structure.

Le chapitre suivant établit un parallèle entre l'évolution de la structure multidimensionnelle et celle des ontologies. La première partie est consacrée à l'étude du domaine des ontologies et à la revue des solutions proposées dans la littérature pour le mapping des ontologies. La seconde partie du chapitre décrit la notion de similarité sémantique et

constitue une synthèse des différents modèles de mesure de similarité sémantique qui sont susceptibles d'être intégrées à une fonction de mapping pour l'établissement de liens sémantiques entre les entités des ontologies.

Dans le chapitre 4, nous présentons notre approche pour le rétablissement de liens sémantiques entre les versions de la structure multidimensionnelle. Nous décrivons le modèle de similarité qui sera employée dans l'élaboration de la fonction de rétablissement de liens sémantiques. Ce modèle de similarité se fonde également sur une phase de segmentation et une phase d'indexation des textes, lesquelles seront décrites dans ce chapitre.

Le chapitre 5 expose l'approche pour la solution au problème de l'évolution géométrique ; différentes méthodes d'indexation de données spatiales y sont décrites, dont la méthode *QuadTree* qui est employée dans notre approche. Nous décrivons comment la représentation commune créée par cette méthode d'indexation permet de constituer des matrices de correspondances sémantiques. Dans une seconde partie, nous présentons l'approche géosémantique qui consiste à intégrer les approches sémantique et géométrique pour permettre le traitement de requêtes spatio-temporelles.

Le chapitre 6 présente l'application que nous avons élaboré pour tester notre approche, les données utilisées, qui proviennent du domaine forestier ainsi que l'analyse des résultats que nous avons obtenus.

Finalement, la conclusion comporte une synthèse des points importants de notre approche et de son originalité comme solution au problème d'évolution de la structure multidimensionnelle, ainsi que les perspectives pour les travaux futurs.

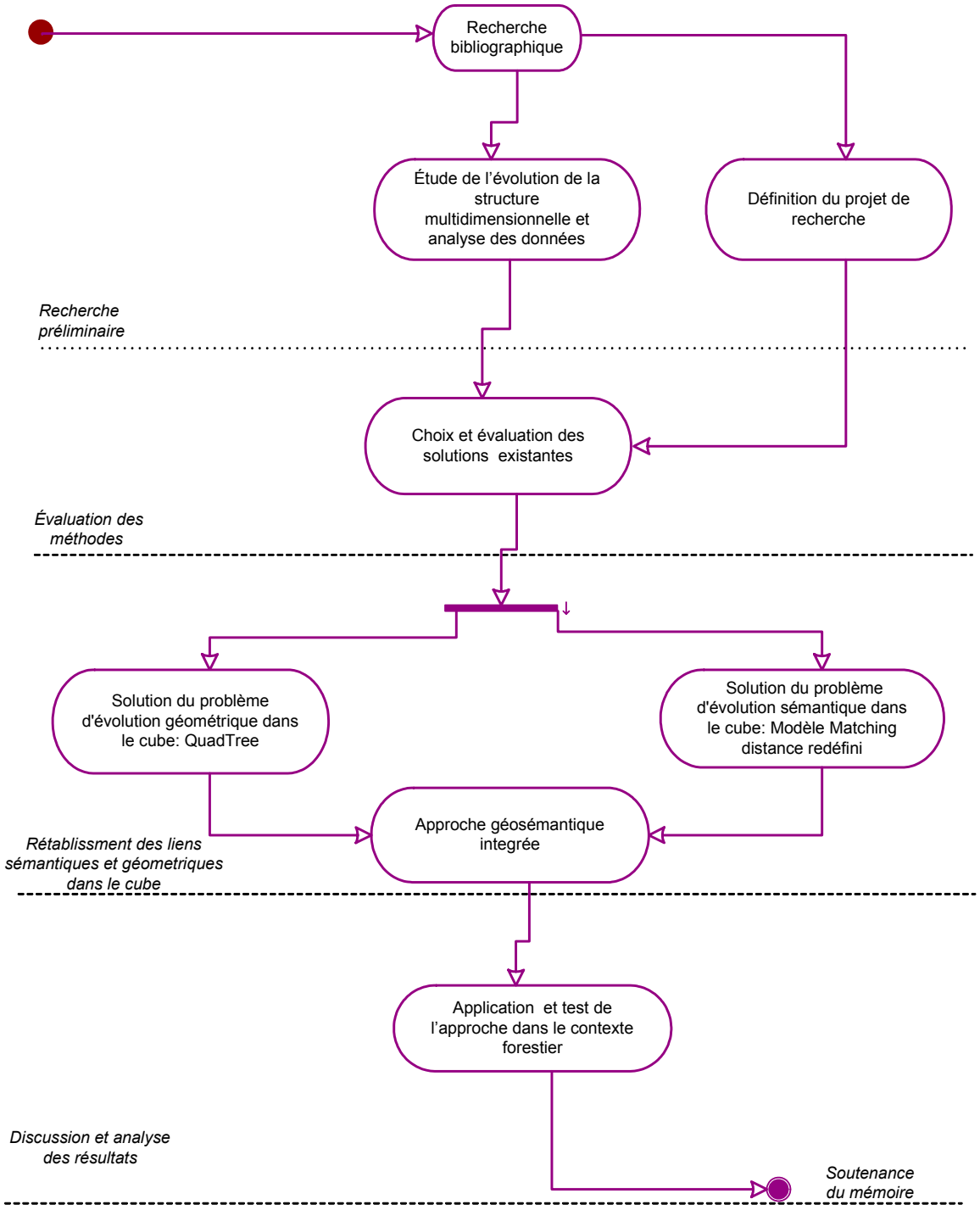

Figure 1.1 : Schéma général de la recherche

Chapitre 2 : Évolution de la structure des bases de données géospatiales multidimensionnelles

2. 1 Introduction

Actuellement, les entreprises ont besoin d'outils et de modèles pour la mise en place de systèmes décisionnels comportant des données évolutives (Pedersen et Jensen, 1999) (Mendelzon et Vaisman, 2000) et permettant de faire des analyses spatiotemporelles. Celles-ci sont devenues primordiales pour que les utilisateurs en situation de prise de décision puissent analyser les données dans un contexte spatial (Zhang et Tsotras, 2001 ; Bédard et al., 2005). Or, les modèles multidimensionnels à la base des systèmes décisionnels possèdent une structure statique qui ne permet pas de prendre en compte les évolutions que la structure multidimensionnelle est amenée à subir, affectant le traitement des requêtes temporelles dans les cubes de données géospatiales. Plusieurs approches ont été proposées comme solution à la problématique de l'évolution de la structure ; bien qu'elles apportent des solutions permettant effectivement à l'utilisateur de mener à bien des analyses temporelles, elles ne sont pas en mesure de traiter tous les types d'évolutions pouvant affecter la structure multidimensionnelle, plus particulièrement les évolutions sémantiques et géométriques de cette dernière.

2. 2 Bases de données géospatiales multidimensionnelles

Les besoins croissants de pouvoir procéder à des analyses sur des données dans le but de prendre des décisions ont motivé le développement de nouveaux systèmes appelés Systèmes d'aide à la Décision (SADs) (Shimm et al, 2002). Les SADs regroupent un ensemble d'informations et d'outils mis à la disposition des utilisateurs pour aider à manière rapide et efficace la prise de décision (Chaudhuri, Dayal 1997). L'architecture d'un SAD comprend des sources de données, à partie desquelles les données sont extraites, transformées et intégrées dans l'entrepôt de données. Ce dernier est une collection de

données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse (Inmon, 1996). Le magasin de données contient une partie des données contenues dans l'entrepôt ; ces données sont orientées-sujet, c'est-à-dire qui sont organisées autour d'un thème d'analyse pertinent pour l'utilisateur. Finalement, le SAD est associé à des outils d'analyse tels que OLAP, et son extension spatiale SOLAP.

Afin de permettre à l'utilisateur d'analyser facilement et intuitivement les données, les entrepôts de données se basent sur une structure multidimensionnelle, qui permet de représenter les données selon plusieurs axes d'analyse, appelés dimensions. Les dimensions peuvent représenter plusieurs thèmes d'intérêt pour l'utilisateur, par exemple l'âge, le lieu géographique, le temps, etc. La structure multidimensionnelle peut être représentée par un cube, que nous illustrons par un exemple issu du domaine forestier à la figure 2.1. Un cube est composé d'éléments appelés cellules. Les cellules contiennent les valeurs d'un fait, habituellement appelées mesures. Les axes du cube correspondent aux dimensions et ils sont gradués par des membres. Dans la structure multidimensionnelle, les dimensions sont des hiérarchies et comportent donc un ensemble de niveaux liés par des relations de classification. La structure d'une dimension est représentée par le schéma de la dimension. Le schéma de la dimension, lorsqu'il est instancié, forme le schéma des instances qui explicite tous les membres de la dimension (figure 2.2).

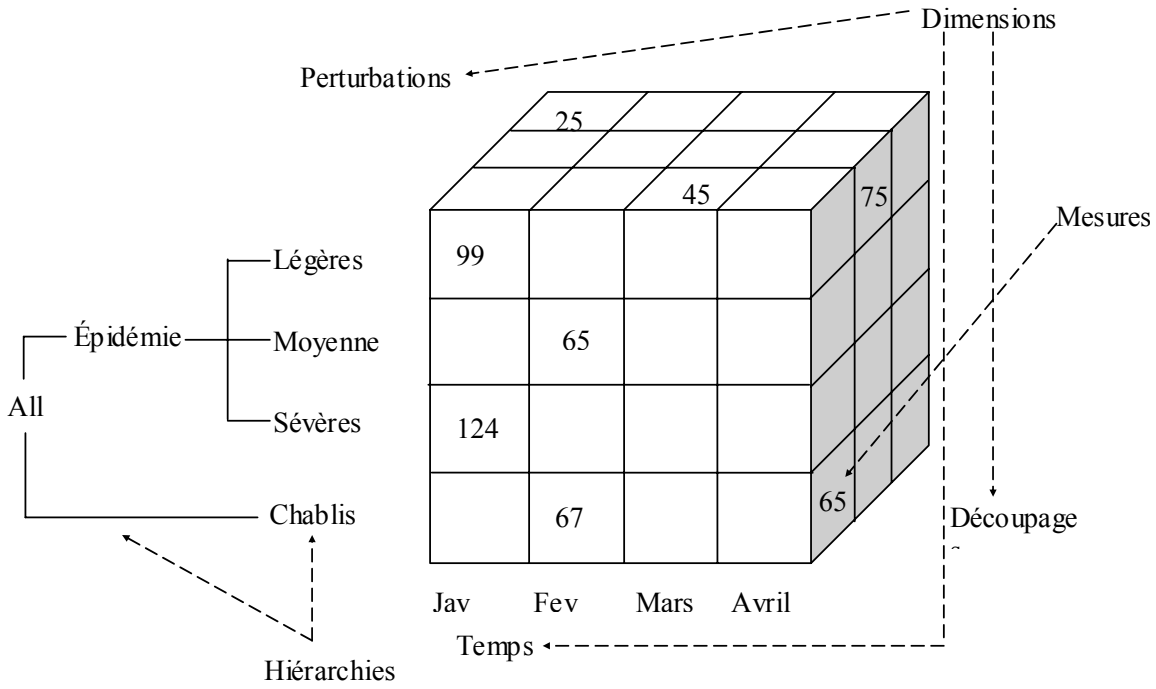

Figure 2.1 : Cube de données multidimensionnel dans le domaine forestier

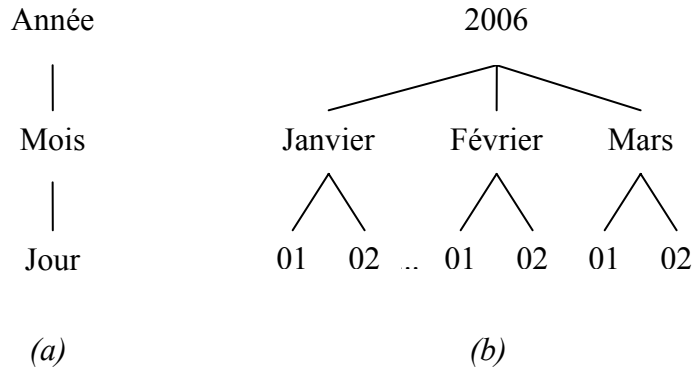

Figure 2.2 : Exemple de schéma de dimension (a) et de schéma des instances de la dimension (b)

Une base de données géospatiales multidimensionnelle comprend également une ou plusieurs dimensions spatiales, dont on distingue trois types (Bédard, 2001): la dimension

spatiale géométrique, non géométrique ou mixte. Dans la dimension spatiale géométrique, tous les niveaux de la hiérarchie sont représentés par des objets géométriques. Par opposition, tous les niveaux de la dimension spatiale non géométrique ne comportent aucune représentation géométrique, bien que les membres soient des entités spatiales. Dans la dimension spatiale mixte, seule une partie des niveaux de la hiérarchie est associée à une représentation géométrique. Dans une dimension spatiale géométrique, les différents niveaux sont liés entre eux par des relations topologiques (Malinowski et Zimányi, 2004) telles que des relations d'inclusion, d'intersection, de rencontre, etc. Comme les dimensions, les mesures peuvent également posséder une composante spatiale (Rivest et al, 2001). Les mesures spatiales peuvent être de trois types, soit des pointeurs spatiaux permettant de lier un objet spatial à sa géométrie, des mesures qui se servent d'opérateurs métriques ou topologiques, ou des mesures combinant les dimensions géométriques composant une hiérarchie donnée (Bédard et al, 2005).

Au niveau conceptuel, la modélisation multidimensionnelle peut se faire soit en étoile, en flocon ou en constellation. Au niveau physique, le cube multidimensionnel peut être implanté selon une approche relationnelle (ROLAP), multidimensionnelle (MOLAP) ou hybride (HOLAP), qui combinent les deux structures; dans ce cas il est stocké selon des rangées multidimensionnelles. Dans ROLAP (Relationnel OLAP), il est implémenté comme un ensemble de tables relationnelles : des tables de dimension et des tables de faits, ces dernières contenant les valeurs des mesures.

Dans les modèles multidimensionnels actuels, les dimensions sont considérées statiques et les faits dynamiques. En réalité, les dimensions doivent également être considérées comme dynamiques puisqu'elles subissent des évolutions au cours du temps. Dans la section suivante, les différents types d'évolution de la structure multidimensionnelle sont présentés.

2. 3 Évolution des bases de données géospatiales multidimensionnelles

Tous les phénomènes susceptibles d'être modélisés et représentés dans une base de données sont appelés à subir des modifications au cours du temps. Par conséquent, les faits qui forment le cube de données multidimensionnelles sont considérés comme dynamiques. La structure des bases de données est également appelée à évoluer au cours du temps, en raison de modifications conceptuelles, de l'expansion des connaissances au sujet des phénomènes modélisés, de l'évolution des classifications, etc. Ainsi, on peut facilement concevoir que dans l'exemple du cube illustré à la figure 2.1, qu'un type de perturbation des peuplements forestiers puisse s'ajouter aux catégories existantes, par exemple les feux de forêt, modifiant ainsi la structure de la dimension *Perturbation*. Ces évolutions deviennent de plus en plus probables puisque les bases de données font partie de systèmes d'information dont la durée de vie tend à s'allonger (Grandi et Mandreoli, 2002).

Au cours de nos recherches, nous avons été amenés à distinguer deux types d'évolution qui peuvent affecter la structure multidimensionnelle, à savoir l'évolution directe et l'évolution indirecte.

Évolution directe

Le besoin de modifier la structure multidimensionnelle se manifeste lorsque des changements extérieurs viennent affecter la réalité modélisée par la base de données. Le concepteur peut, par exemple, choisir de modifier la structure de la base de données pour tenir compte d'une nouvelle représentation du phénomène modélisé, en ajoutant ou en supprimant par exemple des niveaux dans une dimension. L'évolution directe se manifeste lorsqu'un concepteur modifie explicitement la structure de base de données par des opérateurs appropriés de mise à jour. Les méthodes de mise à jour proposent un ensemble d'opérateurs d'évolution permettant, par exemple, d'ajouter ou de supprimer une dimension ou un membre de la dimension. L'action des opérateurs d'évolution sur la structure crée une nouvelle version de cette structure. Les méthodes de mise à jour s'appliquent au niveau conceptuel (Blaschka, 2000 ; McBrien, Poulovassilis, 2002) ou au niveau logique de la structure multidimensionnelle (Hurtado et al., 1999 ; Guerrero, 2002) L'évolution produite

par les méthodes de mise à jour est directe puisque les opérations qui lient une version à l'autre sont connues.

La taxonomie des évolutions de la structure multidimensionnelle présentée dans le tableau 2.1 consiste à énumérer toutes les opérations d'évolutions possibles qui peuvent affecter la structure multidimensionnelle. Les approches qui gèrent l'évolution de la structure, c'est-à-dire qui permettent de créer plusieurs versions, gèrent de manière partielle ou complète ces opérations. Les évolutions possibles peuvent être catégorisées selon qu'elles sont des changements du schéma des dimensions, des changements dans le schéma des instances d'une dimension ou des modifications des faits.

Modification de schéma de dimension	Modification de schéma des instances	Modification de fait
• ajout / suppression d'un niveau • ajout /suppression d'une dimension • ajout / suppression d'un lien hiérarchique • déplacement d'un niveau dans la structure hiérarchique	• ajout / suppression d'un membre • re-classification d'un membre • transformation d'un membre division / fusion d'un membre	• ajout/suppression d'un fait • modification d'un fait

Tableau 2.1 : Taxonomie des évolutions de la structure multidimensionnelle

Évolution indirecte

Le cas de l'évolution indirecte est plus complexe que celui de l'évolution directe puisque ce type d'évolution n'est pas produit intentionnellement en appliquant des opérations d'évolution à la base de données, mais résulte de la modélisation successive d'un même phénomène à plusieurs intervalles de temps. Ces modélisations successives ont pour conséquence de produire plusieurs versions de la structure indépendantes sur le plan de la conception, puisqu'elles sont construites à partir de sources indépendantes, mais liées par la réalité qu'elles représentent. C'est le cas, en particulier, des bases de données qui sont obtenues à partir de la représentation d'un même territoire à différentes époques, comme par exemple les données qui sont recueillies lors des inventaires de la forêt Montmorency. Au fil des inventaires, il est naturel que les données (mesures de surface, de volume

ligneux, etc.) évoluent, mais l'évolution naturelle de la forêt, les interventions humaines, l'évolution des modes d'acquisition, des spécifications et des normes forestières sont également des réalités de la gestion forestière qui affectent l'évolution de la structure multidimensionnelle du cube de données géospatiales. Par exemple, les données de la forêt de Montmorency sont recueillies à partir des inventaires décennaux de 1973, 1984, 1992 et 2001. À chaque inventaire correspond une version de la structure (un cube de données géospatiales) ayant sa propre structure. Pourtant, ces versions sont liées car elles représentent toutes la forêt Montmorency mais à différentes époques (figure 2.3).

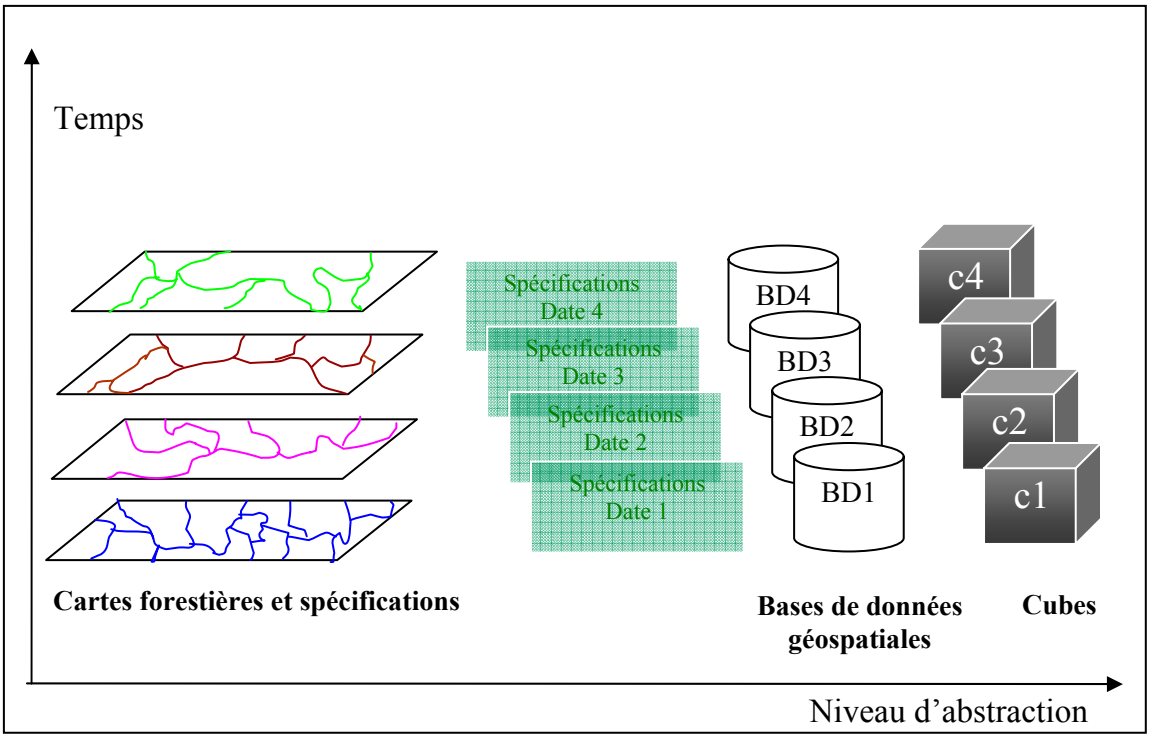

Figure 2. 3: Propagation de l'évolution aux cubes de données géospatiales.

En général, les types d'évolutions pouvant affecter la structure multidimensionnelle se produisent au niveau de la réalité physique ou de la réalité institutionnelle. La réalité physique concerne les phénomènes qui sont indépendant de l'humain alors que la réalité institutionnelle concerne tous les phénomènes qui sont définis par ce dernier sur la réalité physique (Bittner, 2001). Dans les bases de données géospatiales, l'évolution comparative de la structure multidimensionnelle peut être causée par une évolution au niveau de la

couverture du territoire modélisé, au niveau de la cueillette de données ou au niveau de la représentation conceptuelle, ou spécifications de conceptualisation (ontologie). La figure 2.4 résume les facteurs qui peuvent influencer la couverture du territoire.

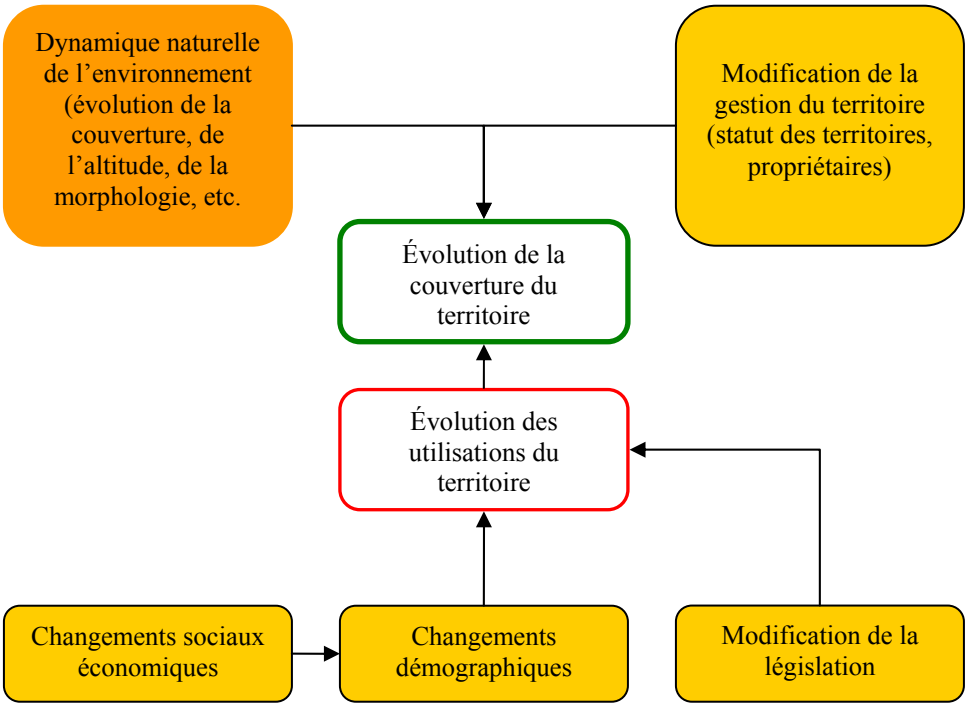

Figure 2.4 : Facteurs à la source de l'évolution de la couverture du territoire

L'évolution de la couverture du territoire a pour conséquence que les tracés territoriaux (par exemple les frontières des peuplements forestiers déterminées à partir des photos aériennes) vont également évoluer selon une nouvelle réalité ; par conséquent, la dimension spatiale va également évoluer, des peuplements pouvant disparaître ou être ajoutés, transformés, re-classifiés, etc.

Au niveau de la cueillette de données, l'évolution des modes et des outils d'acquisition des données, la modification des normes, de classification et des spécifications peuvent être à la source de l'évolution de structure multidimensionnelle. Par exemple, la classification des zones forestières par intervalle de densité peut être modifiée : dans un premier inventaire forestier, les zones présentant une couverture se situant entre 10% et 25% peuvent être classifiées comme *peu denses* alors que dans un second inventaire, des zones seraient

classifiées sous la catégories *peu denses* si elles présentent une couverture se situant entre 15% et 35%.

Finalement, l'évolution de la spécification des conceptualisations (évolution de l'ontologie) se répercute également sur la structure multidimensionnelle. Par exemple, le concepteur peut décider qu'il est plus judicieux de classifier le groupement d'essences *Résineux et peupliers* dans la catégorie *Essences mélangées* plutôt que dans la catégorie *Résineux*, ce qui devrait se traduire par une reclassification d'un membre de la dimension *Essence*.

L'évolution des réalités physique et institutionnelle cause l'évolution sémantique et géométrique de la structure multidimensionnelle. L'évolution géométrique se manifeste exclusivement au niveau de la dimension spatiale géométrique lorsque les entités géométriques des membres de celle-ci sont modifiées (redéfinition des frontières, changement de position, de forme, etc.). L'évolution sémantique peut affecter à la fois les dimensions spatiales ou les dimensions non spatiales lorsque la définition des membres est modifiée où qu'un membre est reclassifié dans un autre niveau de la hiérarchie d'une dimension. Dans la section suivante, nous examinons les impacts de l'évolution (directe et indirecte) sur la structure multidimensionnelle.

2. 4 Impacts de l'évolution des bases de données géospatiales multidimensionnelle

Un des objectifs premiers de l'analyse étant de pouvoir étudier les données historiques afin de pouvoir observer les variations, tendances et corrélations dans l'ensemble des données. Le temps constitue un élément clef pour assurer l'efficacité des modèles de données multidimensionnels (Pedersen, Jensen, 1999). Les entrepôts de données permettent de stocker les données en tenant compte du temps, soit en soutirant les attributs de temps de la base de données opérationnelle, ou en ajoutant un attribut de temps lors des processus de mise à jour ; donc les faits, parce qu'ils possèdent cet attribut temporel, peuvent être analysés chronologiquement. Il en va autrement pour les dimensions, qui sont considérées comme statiques. Dans la réalité, il est fréquent que le schéma d'une dimension ou que le schéma des instances d'une dimension doivent être modifiés, comme nous l'avons montré dans la section précédente.

Afin de pallier à ce problème et de permettre aux évolutions qui se produisent dans la réalité d'être représentés dans les bases de données, des modèles supportant l'évolution et d'autres gérant le versioning ont été proposés, tant au niveau du schéma conceptuel qu'au niveau de la structure multidimensionnelle. Au niveau du schéma conceptuel, une base de données supporte l'évolution de schéma si elle permet de modifier un schéma sans perdre de données et elle gère le versioning de schéma si elle permet de faire des requêtes sur toutes les données par le biais d'une version quelconque définie par l'utilisateur (Jensen et al.,1998). Dans ce mémoire, nous considérons que cette définition peut se traduire de manière équivalente pour la structure multidimensionnelle, c'est-à-dire qu'une base de données multidimensionnelle supporte l'évolution de structure si elle permet de modifier la structure sans perdre de données et elle gère le versioning de la structure si elle permet de faire des requêtes sur toutes les données par le biais d'une version quelconque définie par l'utilisateur. Comme il en fut discuté dans la section précédente, les approches pour l'évolution de structure proposent des opérateurs pour modifier la structure de la base de données (c.f. tableau 2.1), et les approches de versioning permettent de conserver toutes les versions créées pour les rendre accessibles à l'utilisateur. Ces approches pour l'évolution de structure ont pour effet de produire une évolution directe de la structure multidimensionnelle.

Une des conséquences de l'évolution directe de la structure des dimensions est que les faits peuvent être représentés dans différentes structures résultants de cette évolution et, par conséquent, des requêtes exécutées à différents points dans le temps peuvent fournir des résultats différents et parfois contradictoires (Thurnheer, 2000; Body et al, 2003; Bebel et al, 2004). De plus, on peut répondre à une requête temporelle en représentant les données de la réponse dans une des versions de la structure ou en temps consistant (Body et al, 2002), c'est-à-dire en représentant les données dans leur structure initiale respective, telles qu'elles ont été recueillies. Par conséquent, les réponses obtenues ne sont pas nécessairement cohérentes et pour un utilisateur non informé de l'évolution des dimensions, cela peut mener à une interprétation inexacte des résultats.

Le fait qu'il existe plusieurs représentations structurelles empêche également d'agréger les données et de les comparer dans le temps. De plus, même en agrégeant les réponses

obtenues, nous obtiendrions des valeurs dont l'interprétation peut mener à une conclusion erronée de la part de l'utilisateur.

L'évolution indirecte de la structure multidimensionnelle a pour conséquence de créer des versions de la structure qui n'ont pas de liens entre elles, puisqu'elle discrétise un phénomène temporel continu sans conserver les liens qui unissent les objets modélisés. Ce type d'évolution est donc plus complexe à traiter que le cas de l'évolution directe où les liens entre les objets sont généralement connus. En effet, le fait par exemple de reclassifier un membre de la structure (en redéfinissant les relations de ce membre avec les autres membres de la hiérarchie d'une dimension) implique que l'on peut, au moment de cette opération, conserver (par des moyens suggérés dans plusieurs approches que nous verrons à la section suivante) l'information sur cette évolution.

Concrètement, l'évolution indirecte de la structure multidimensionnelle se traduit par des impacts, comme dans le cas de l'évolution directe, sur la réponse aux requêtes temporelles. L'évolution indirecte de la dimension spatiale, qui se manifeste lorsque les frontières des entités spatiales, telles que des peuplements forestiers, évolue, se traduit par l'impossibilité de répondre à des requêtes spatiotemporelles sur les entités spatiales puisque ces dernières ne sont pas liées dans le temps dans les différentes versions de la base de données. Par exemple, considérons la requête suivante : quelle a été la densité de chaque peuplement au cours des vingt dernières années ? Les peuplements étant des entités spatiales qui existent seulement pour un des inventaires forestiers, cette requête ne peut être répondue puisque les peuplements appartenant à des époques distinctes ne sont pas liés entre eux. L'évolution des dimensions non spatiales peut fausser la réponse aux requêtes temporelle, par exemple si la définition d'un membre a été modifiée d'une version à l'autre. Par exemple, la dimension *Âge* du modèle multidimensionnel de la forêt de Montmorency qui possède une structure différente d'un inventaire à l'autre puisque les classes d'âge sont associées à des intervalles d'âge différents (figure 2.5).

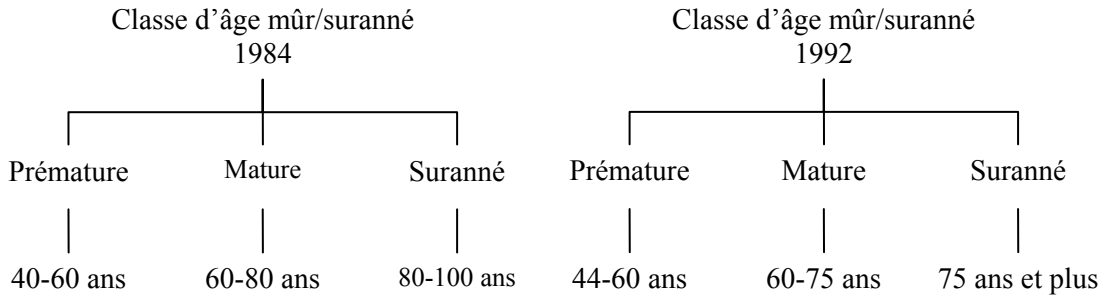

Figure 2.5 : Hiérarchies partielles de la dimension Âge du modèle multidimensionnel de la forêt de Montmorency entre 1984 et 1992 (Rebout et al., 1998)

Une requête sur le volume ligneux par classe d'âge recevrait une réponse inconsistante puisque par exemple pour la classe d'âge *suranné* l'utilisateur obtiendrait pour l'inventaire de 1984 la somme du volume ligneux pour les arbres dont l'âge se situe entre 80 et cent ans alors que pour l'inventaire de 1992, il obtiendrait la somme du volume ligneux pour tous les arbres dont l'âge est supérieur à 75 ans. Ces réponses n'étant pas comparables, l'utilisateur peut en déduire des conclusions erronées en ce qui concerne l'évolution du volume ligneux de la classe d'âge *suranné*. La problématique engendrée par l'évolution de la structure multidimensionnelle, qu'elle soit directe ou indirecte, est que l'aspect dynamique des phénomènes modélisés complique le traitement des requêtes spatiotemporelles. L'impossibilité de traiter les réponses temporelles et les réponses erronées affecte directement le processus de prise de décision qui doit alors se baser sur de données incomplètes ou faussées. Dans la section suivante les différentes méthodes actuelles pour résoudre le problème d'évolution de la structure multidimensionnelle et leurs points forts et faibles sont identifiés.

2. 5 Solutions existantes pour le problème d'évolution de la structure multidimensionnelle

Dans cette section, nous présentons les approches principales qui ont été proposées pour résoudre la problématique de l'évolution de la structure multidimensionnelle pour permettre de traiter les requêtes temporelles. Le premier type d'approche est de permettre de concrétiser les modifications des dimensions par l'utilisation d'opérateurs; ces

approches de mises à jour ont été brièvement discutées dans les sections précédentes notamment comme étant à la source de l'évolution directe de la structure. Actuellement, il est reconnu que les approches de mise à jour ne peuvent suffire pour prendre en compte l'évolution de la structure et n'offrent aucune solution pour traiter efficacement les requêtes temporelles (Body, Bédard et al., 2003) Les approches proposées plus récemment (Vaisman, 2001) incorporent souvent leur propre approche de gestion de l'évolution et se concentrent sur la problématique des requêtes temporelles. Celles-ci peuvent être considérées au niveau conceptuel, soit les modèles temporels, ou au niveau de l'implantation, par les méthodes de versioning.

2.5.1 Modèles temporels

La dimension temporelle est généralement considérée comme une dimension quelconque dans les bases de données conventionnelles, de telle sorte que la dimension temporelle, orthogonale aux autres dimensions, ne peut être utilisée pour représenter les changements dans la structure des dimensions (Eder, Koncilia, Morzy, 2002). Cette approche ne permet pas d'exprimer des requêtes temporelles complexes comparables aux possibilités offertes dans les bases de données temporelles (Teste, 2000) où l'aspect temporel caractérise non seulement les faits, selon l'approche traditionnelle, mais également les dimensions. Au niveau conceptuel, la solution du modèle temporel est d'ajouter un aspect temporel aux composantes du modèle multidimensionnel de base. Le concept de temps valide (Snodgrass, 1995), qui représente l'intervalle de temps pendant lequel une entité est vérifiée dans la réalité (par opposition au temps transactionnel qui vérifie l'intervalle de temps pendant lequel une entité existe dans la base de données), est un concept central de ce type d'approche. Les modèles temporels associent un intervalle de temps valide aux relations hiérarchiques, niveaux et membres du schéma de la dimension et/ou du schéma des instances (Mendelzon, Vaisman, 2000; Vaisman, 2001; Pedersen et al, 2001). Certains modèles incluent, en plus du temps valide, d'autres aspects temporels, tel que le temps transactionnel (Koncilia, 2003) et l'instant où les données ont été chargées dans l'entrepôt de données (Data Warehouse Loading Time, DWLD, généré par l'entrepôt de données) (Malinowski, Zimanyi, 2006), considérant que des aspects temporels plus variés sont susceptibles de fournir de l'information utile pour l'utilisateur, par exemple pour procéder à

des analyses de traçabilité. Certains modèles définissent leur propre langage pour exprimer des requêtes temporelles (Vaisman, 2001 ; Terenziani et al, 2006). Les modèles temporels plus conventionnel s'appuient sur une représentation discrétisée du temps, alors que d'autres modèles en offrent une représentation continue (Erwig et al, 1999 ; Ahmed et al, 2004), permettant ainsi de représenter des phénomènes continus dans le temps.

2.5.2 Méthodes de versioning

Le principe du versioning, à l'opposé de celui d'évolution de schéma (méthode de mise à jour), consiste à suivre l'historique du schéma en conservant toutes les versions de la structure. Les méthodes de versioning représentent la solution du modèle temporel au niveau de l'implantation. Les bases de données multiversions intègrent donc des concepts des bases de données temporelles, comme le concept de temps valide et se distinguent des bases de données transactionnelles qui permettent de visualiser les données uniquement dans leur représentation actuelle. Une version possède un intervalle de temps valide déterminé par l'intervalle de temps pendant lequel tous les éléments (niveaux de dimensions, membres du schéma des instances, relations hiérarchiques entre les membres, etc.) sont valides. Quelques solutions offrent à l'utilisateur des versions alternatives (Balmin et al., 2000 ; Eder et Koncilia, 2002), à partir desquelles l'utilisateur peut analyser des scénarios fictifs (*analyse what-if*) en créant un structure virtuelle. Le schéma de l'entrepôt de données peut être représenté par un graphe, composé de nœuds et d'arcs, sur lequel des opérations d'évolution de schéma permettent de créer un historique des versions de schémas (Golfarelli et al., 2004); de plus, un schéma augmenté, qui fait le lien entre les versions, est créé afin de permettre de répondre à des requêtes qui portent sur plusieurs versions, l'utilisateur pouvant choisir une version quelconque pour les requêtes.

Cependant, le principe de versioning, à lui seul, n'est pas suffisant pour traiter adéquatement les requêtes temporelles, puisqu'il est nécessaire non seulement de connaître l'état de la structure à une certaine époque mais également de pouvoir établir des liens entre les versions. Les méthodes plus adaptées de versioning intègrent des fonctions de mapping dont la fonction est de lier deux versions de la structure.

L'approche de Body (Body et al., 2002) se situe également au niveau conceptuel et intègre le concept de temps valide pour les membres et liens hiérarchiques. Les évolutions sont considérées uniquement au niveau des instances, puisqu'une évolution au niveau du schéma de la dimension peut se traduire par une série d'évolutions au niveau du schéma des instances; cette approche permet entre autre de supporter les hiérarchies complexes. L'utilisateur peut choisir une version quelconque pour les requêtes ou obtenir une réponse en temps consistant, c'est-à-dire en représentant les données issues de différentes versions dans leur structure initiale. Cette approche se distingue par la possibilité d'établir des relations de mapping entre les versions, accompagnées d'un indice de confiance indiquant leur qualité. Cependant, puisqu'une seule table de fait centrale est employée pour stocker toutes les versions des données, seuls les changements sur le schéma de dimension et schéma des instances sont supportés.

Une approche originale (Eder, Koncilia, 2002) intègre également des fonctions de mapping entre les différentes versions, lesquelles sont stockées sous forme de matrices et utilisées pour transformer les mesures associées à une version vers la structure d'une autre version, permettant de traiter l'évolution au niveau du schéma de la dimension et du schéma des instances. L'approche des entrepôts multiversions (Morzy, Wrembel, 2004), gère dans un seul entrepôt plusieurs versions de la structure formant un graphe de dérivation, où les nœuds représentent des versions, et les arcs lient deux versions adjacentes dans le temps ou des versions alternatives (créées à partir de versions réelles à des fins de simulation).

Nous présentons un tableau récapitulatif (tableau 2.2) qui permet de comparer les solutions exposées selon le type d'évolution qu'elles traitent, leurs apports et leurs désavantages. Les modèles de mise à jour proposent des opérateurs permettant de modifier la structure, soit au niveau du schéma de la dimension et des faits (Blaschka, 2000) ou au niveau, également, du schéma des instances des dimensions (Hurtado et al, 1999). Nous n'avons pas discuté de manière approfondie de ces approches puisqu'il est admis que les approches de mise à jour sont conçues pour concrétiser l'évolution d'une réalité dans la base de données. Cependant, les approches de mise à jour ne constituent pas une solution à la problématique de l'évolution de la structure multidimensionnelle puisqu'elles ne conservent pas les états antérieurs de la structure et par conséquent ne permettent pas de retracer l'évolution. Les

approches de mise à jour peuvent constituer une solution à la problématique de l'évolution uniquement si elles sont combinées à des approches permettant de conserver les différentes versions de la structure créée et de conserver des liens entre elles. Parmi les approches des modèles temporels, nous avons représentés la seule approche la plus complète, c'est-à-dire celle qui gère les évolutions à la fois au niveau du schéma de la dimensions et au niveau du schéma des instances des dimensions (Mendelzon, Vaisman, 2000). Néanmoins, nous considérons que les approches des modèles temporels ne peuvent être complètes puisque, même si elles permettent de conserver la durée de validité des membres de la structure multidimensionnelle, elles ne permettent pas de traiter les requêtes temporelles (des requêtes qui portent sur plusieurs versions de la structure) en raison de l'absence de liens entre les membres. L'approche de versioning proposées par Kimball (Kimball, 1996) permet d'identifier différentes versions de la structure multidimensionnelle mais ne supporte que certains types d'évolution au niveau du schéma des instances (modification, ajout et suppression de membres). Les trois dernières approches présentées semblent les plus intéressantes car elles sont les seules qui permettent réellement de traiter des requêtes temporelles sur des membres qui ont évolué en raison de la présence de fonction de mapping entre les membres du schéma des instances. Dans l'approche de Eder et Koncilia, les fonctions de mapping entre les membres des schéma des instances des dimensions sont stockées dans des matrices qui permettent de transformer, lors de la requête, les mesures du cube dans une version de la structure choisie par l'utilisateur et ce selon une ou plusieurs dimensions. Ceci offre plus de flexibilité que les deux autres approches, où les mesures devant être représentées selon différentes structures sont stockées dans une table de faits multiversions. Cependant, parmi toutes les approches présentées, aucune ne considère le cas où les liens entre les versions sont inconnus et doivent être préalablement identifiés, c'est-à-dire que l'on considère que les fonctions de mapping qui lient les membres doivent être connues et intégrées lors de la mise à jour de la structure. Certaines approches de versioning (Eder et Koncilia, 2001; Body et al, 2002 ; Morzy et al, 2004) permettent d'intégrer des fonctions de mapping qui tiennent compte des relations topologiques (du taux d'inclusion, plus précisément, qui mesure le degré de superposition entre deux entités géométriques) entre des membres du schéma des instances de la dimension spatiale de différentes versions du cube, mais dans aucun cas elles ne considèrent comment ces

relations topologiques peuvent être dérivées ; autrement dit, elles doivent être identifiées d'avance. De plus, les solutions proposées ne tiennent pas compte de l'évolution sémantique et géométrique de manière simultanée.

Type d'approche		Type d'évolution supporté	Apports	Désavantages
Mises à jour	(Blaschka, 2000)	-Schéma de dimension -Faits	Mise à jour complète du schéma de dimension et des faits	Pas d'historique et perte d'information
	(Hurtado et al, 1999)	-Schéma de dimension -Schéma des instances	Maintenances des vues	Pas d'historique et perte d'information
Modèles temporels	(Mendelzon, Vaisman, 2000)	-Schéma de dimension -Schéma des instances	-Méta-requêtes -Résultats des requêtes en temps consistant ou dans la dernière version -liens entre les versions lors de fusions et divisions de membres	-Utilisateur ne peut choisir les versions antérieures -Pas de liens entre les membres de la structure
Versioning	(Kimball, 1996)	-Schéma des instances (partiellement supporté)	Résultats des requêtes en temps consistant	Altération possible des résultats
	(Eder, Koncilia, 2001)	-Schéma des instances	-Utilisateur peut choisir entre toutes les versions -Mapping entre les versions	Pas de résultats des requêtes en temps consistant
	(Body et al, 2002)	-Schéma de dimension -Schéma des instances -Mesures	-Mapping entre les versions -Traitement des structures hiérarchiques complexes -informe l'utilisateur de l'évolution	-Une seule table de faits statique
	(Morzy et al, 2004)	-Schéma de dimension -Schéma des instances -Faits	-Versions alternatives -informe l'utilisateur de l'évolution	-Mapping partiels entre les versions

Tableau 2.2: Comparaison des différentes solutions au problème d'évolution de la structure multidimensionnelle

2. 6 Conclusion

Dans ce chapitre, nous avons présenté les notions fondamentales liées à la structure multidimensionnelle des cubes de données géospatiales, ainsi que les causes et différentes formes d'évolution de la structure. Nous avons constaté que l'évolution peut se manifester lorsque des opérations de mise à jour, telles que l'ajout ou la suppression de membres du schéma des instances, ou lorsqu'une même réalité est représentée à différentes époques par différents cubes de données géospatiales. De plus, l'évolution ne se limite pas à des ajouts, suppressions, fusions de membres, etc., mais peut également être sémantique, par exemple si la définition des membres du schéma des instances de la dimension est modifiée, et géométrique, par exemple si les membres du schéma des instances de la dimension spatiale voient leur géométrie changée. Dans ces cas d'évolution, les liens entre les différents cubes ne sont pas conservés, et les résultats aux requêtes spatio-temporelles sont faussés ou impossibles à obtenir. Par conséquent, toute solution à la problématique de ce type d'évolution doit proposer une méthode de rétablissement de liens avant de pouvoir permettre de traiter les requêtes temporelles. Notre approche comporte donc deux volets qui seront présentés dans les chapitres 4 et 5, soit l'approche de rétablissement de liens sémantiques et l'approche de rétablissement de liens géométriques. Avant de présenter ces approches, dans le chapitre suivant, nous abordons le domaine des ontologies et de la similarité sémantique qui nous a permis de développer la méthode de rétablissement de liens sémantiques.

Chapitre 3 : Ontologies et similarité sémantique

3. 1 Introduction

Les approches présentées dans le chapitre précédent pour la problématique de l'évolution de la structure multidimensionnelle ont montrées que bien que la structure puisse subir des évolutions sur le plan sémantique et géométrique, ces types d'évolution ne sont pas considérés dans les solutions proposées. Ces solutions se focalisent plutôt sur la gestion des évolutions génériques (ajout, suppression, reclassification de membres, etc.) mais pas sur des évolutions sémantiques plus complexes telles que, par exemple, la redéfinition des membres du schéma des instances des dimensions qui affecte les cubes de données forestières. Le problème de l'hétérogénéité et de l'évolution sémantique est, par contre, largement documenté dans le domaine des ontologies puisque ces dernières se rattachent précisément à la sémantique des données. Plusieurs parallèles peuvent être établis entre l'évolution de la structure multidimensionnelle et celle de l'ontologie ; par exemple, les types d'évolution identifiés dans le modèle multidimensionnel trouvent leur analogue dans les modifications des ontologies. L'approche que nous proposons pour le rétablissement des liens sémantiques entre les membres de la structure multidimensionnelle s'inspire des approches de mapping entre les ontologies qui utilisent un modèle de similarité sémantique pour établir les relations entre les concepts. Nous proposons donc dans ce chapitre d'exposer quelques solutions qui ont été mises en œuvre ainsi qu'une revue des différents types de mesure de similarité qui peuvent être employées dans l'élaboration des fonctions de mapping. La première section de ce chapitre traite des concepts fondamentaux liés aux ontologies, puis de la problématique de l'hétérogénéité et de l'évolution de ces dernières. Nous y présentons également les solutions proposées pour ces problématiques, en mettant l'accent sur les solutions de mapping entre les ontologies. Dans la seconde section, nous présentons les notions de base sur le concept de similarité sémantique et une revue des différents modèles de similarité sémantiques qui sont susceptibles de nous intéresser pour constituer, dans notre contexte, un modèle de similarité sémantique adapté pour le rétablissement des liens entre les membres de la structure multidimensionnelle.

3. 2 Les ontologies

Les ontologies constituent une théorie de la connaissance qui permet de formaliser et de synthétiser les connaissances issues d'un domaine particulier (Chandrasekaran et al, 1999). Avec l'accès croissant à plusieurs ontologies indépendantes, la problématique de la discordance entre les ontologies est devenue un sujet de plus en plus documenté en intelligence artificielle. Cette discordance apparaît aussi, comme dans le cas de la structure multidimensionnelle, quand l'ontologie évolue. Les solutions de mapping peuvent être envisagées pour résoudre le problème de l'hétérogénéité ou le problème de l'évolution des ontologies. Afin de pouvoir discuter de ces solutions, nous débutons par l'exposé de quelques notions de base utiles à la compréhension de ce domaine.

3.2.1 Définition de l'ontologie

Dans le domaine de l'intelligence artificielle, on définit l'ontologie comme une spécification explicite d'une conceptualisation (Gruber, 1995), autrement dit une ontologie est constituée par un vocabulaire spécifique utilisé pour décrire une certaine réalité, en plus d'un ensemble propositions explicitant la signification sous-tendue par ce vocabulaire (Guarino, 1998). Elle peut prendre la forme d'un thésaurus, réseaux sémantiques, taxonomie, modèle conceptuel, répertoire de données, etc. (Brodeur, 2004) Les ontologies sont utilisées pour la représentation des connaissances, la recherche, l'extraction, l'intégration et le partage d'information, etc. L'ontologie exprime donc des connaissances qui sont validées par une communauté donnée.

Dans les approches les plus formelles, une ontologie est une théorie logique qui vise à représenter un domaine de discours en terme de concepts et de relations entre les concepts formant un graphe et d'axiomes définissant des affirmations conceptuelles qui déterminent les contraintes sur les concepts, les relations et les instances qui représentent les entités concrètes du domaine (Gomez-Perez, 1999). Dans ce cas, l'ensemble des concepts et des relations qui forment un graphe est appelé la signature ontologique et l'ensemble des contraintes sur les entités de l'ontologie est appelé l'ensemble des axiomes ontologiques (Kalfoglou et Schorlemmer, 2003).

Dans le cas le plus simple, une ontologie est une structure hiérarchique (taxonomie) ou les concepts sont liés par des relations de généralisation et de spécialisation. Les concepts y sont souvent désignés par les nœuds et les relations par des arcs. Les concepts peuvent être caractérisés par des attributs dont les valeurs sont de différents types : nombres, booléens, intervalles, mots, etc. La forme des différentes ontologies est conditionnée par le rôle qui leur est désigné. Ce rôle permet de désigner plusieurs types d'ontologies présentées dans la section suivante.

3.2.2 Types d'ontologies

Les éléments de l'ontologie (concepts, relations, axiomes et instances) sont définis explicitement avec un langage dont la sémantique est plus ou moins formelle, dépendant du degré d'abstraction voulu. Par conséquent, les ontologies présentent des degrés d'abstraction différents dépendamment du domaine auquel elles sont dédiées. On identifie trois types d'ontologies selon un niveau décroissant d'abstraction (figure 3.1) : les ontologies globales (*Top-Level Ontology*) les ontologies de domaine, ou dédiées à une tâche plus spécifique, et les ontologies d'application (Guarino, 1998). Les ontologies globales, c'est-à-dire celles qui présentent le plus haut niveau d'abstraction et de généralité, sont les ontologies formelles, car elles sont issues d'un développement systématique, rigoureux et axiomatique de la logique de toutes les formes et modes d'existence. L'adoption de principes rigoureux dans la conception de l'ontologie formelle répond au besoin de disposer de connaissances pouvant être partagées et transférées d'un contexte à l'autre. Elles sont dédiées à des utilisations générales (ex : WordNet). Une ontologie formelle est donc une théorie des distinctions formelles entre les éléments d'un domaine, indépendamment de leur réalité (Guarino, 1997). Au second degré d'abstraction, on trouve les ontologies de domaine qui sont limitées à la représentation de concepts dans des domaines donnés (géographie, médecine, écologie, etc.) et qui spécialise les concepts de l'ontologie globale. Finalement, les ontologies d'applications offrent le plus fin niveau de spécificité, c'est-à-dire qu'elles sont dédiées à un champ d'application précis à l'intérieur d'un domaine et décrivent le rôle particulier des entités de l'ontologie de domaine dans ce champ. Par exemple, l'ensemble des spécifications sur la forêt de Montmorency constitue

une ontologie d'application qui spécifie les concepts généraux pouvant provenir d'une ontologie de domaine forestier générale.

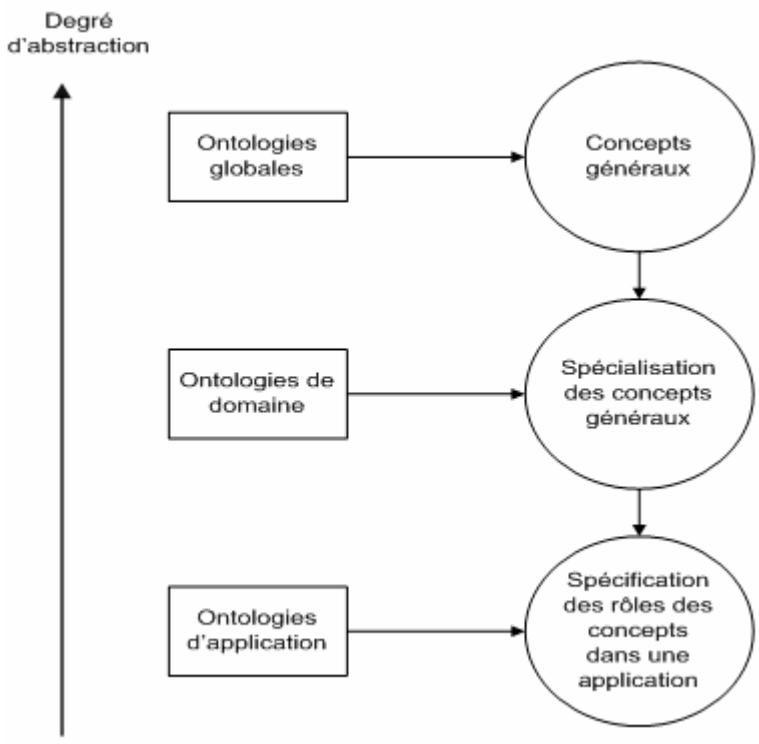

Figure 3.1 : Types d'ontologies

3.2.3 Ontologies spatiales

Les concepts qui composent une ontologie peuvent être de diverse nature, en particulier elles peuvent être des entités spatiales ; dans ce cas l'ontologie elle-même est dite *ontologie spatiale*. Les entités spatiales peuvent être particulièrement importantes dans une ontologie, en particulier celles décrivant les primitives géométriques qui sont décrite par des normes ISO/TC211, OGC, les technologies SIG, etc. et elles viennent enrichir l'aspect sémantique. Dans le cadre de cette recherche, nous ne traitons pas la problématique de l'hétérogénéité de ces normes. De plus, nous ne traitons pas la sémantique de la géométrie des différentes spécifications d'acquisition de données (ex : données topologique). Cette problématique ne fait pas partie des objectifs de notre recherche.

Dans cette section, les aspects principaux de la représentation des entités spatiales dans les ontologies sont présentés.

Les entités spatiales constituent généralement des objets de référence décrits par un ensemble d'attributs descriptifs et une spatialité c'est-à-dire des attributs spatiaux et des relations spatiales. Les relations spatiales comprennent les relations topologiques d'inclusion, de superposition, de contiguïté, etc. ainsi que les relations de position relative (au-dessous, au-dessus, parallèle, perpendiculaire, etc.). Par exemple, les unités de base du découpage du territoire forestier, les peuplements, sont liées par des relations d'inclusions avec les unités de découpage composées qui sont formées de plusieurs unités de base. Les attributs spatiaux incluent les dimensions des entités dans le système de référence : hauteur, aire, orientation, sens, etc., la forme des entités (lignes, point ou polygones) et le système de référence.

Une autre caractéristique des ontologies spatiales est que les données spatiales qui font référence à un sujet particulier y sont souvent regroupées par thème, ou micro-théorie; comme par exemple dans une carte géographique nous pouvons identifier des zones géologiques, administratives, touristiques, etc. Le thème définit un contexte particulier de l'ontologie ; il est donc constitué d'un regroupement d'entités et de relations, une entité pouvant faire partie de plusieurs thèmes. Les thèmes créent des domaines naturels pour guider l'utilisateur dans la recherche d'information ; ils peuvent être utilisés dans les modèles de mesure de similarité sémantique qui tiennent compte du contexte (Zurita, 2004). Compte tenu de cette définition, l'ontologie spatiale peut être considérée comme l'union de plusieurs thèmes.

Tous les types d'ontologies présentés précédemment sont actuellement liés par une même problématique : celle de leur hétérogénéité et de leur évolution qui est présentée dans la section suivante.

3.2.4 Problématique de l'hétérogénéité et de l'évolution des ontologies

Bien que l'ontologie vise à constituer une représentation du monde réel qui puisse être acceptée par tous les membres d'une communauté, les possibilités de représentation sont

variables et par conséquent les ontologies diffèrent malgré les efforts de normalisation. De plus, la normalisation n'est pas toujours souhaitable car la spécificité d'un domaine par rapport à un autre tient en partie à une représentation et à une conception différente des connaissances. Les différentes conceptualisations engendrent l'hétérogénéité des ontologies, une problématique de plus en plus explorée en raison de la multiplication de leur nombre et l'augmentation de leur accessibilité. Ces hétérogénéités limitent les possibilités d'interopérabilité entre les ontologies.

L'hétérogénéité peut se manifester sur deux niveaux : au niveau du langage ou au niveau ontologique (Klein, 2001). Les hétérogénéités au niveau du langage se manifestent quand le langage de spécification diffère entre deux ontologies : les classes, les relations ne sont pas définies de la même manière. Les hétérogénéités de langage se situent sur le plan de la syntaxe, de la représentation logique, de la sémantique du langage ou de l'expressivité. La syntaxe définit la structure des représentations ; des différences sur le plan de la syntaxe concernent le formalisme plutôt que le contenu. Deux ontologies qui diffèrent sur le plan de la syntaxe ne peuvent échanger des données car celles-ci ont des formats différents. Des ontologies peuvent différer sur le plan des représentations logiques, c'est-à-dire que les notions logiques sont exprimées de manière différente, par exemple si dans une ontologie, le fait que deux concepts A et B soient non disjoints s'exprime par A intersect (B) = true et que dans une autre ontologie cette même relation s'exprime par $A \cap B \neq 0$. La sémantique du langage est différente si deux éléments identiques des ontologies ont une signification différente, c'est le cas des termes polysémiques. Finalement des différences sur le plan de l'expressivité impliquent que des ontologies peuvent exprimer certaines notions alors que d'autres ne le peuvent pas, par exemple si une ontologie peut exprimer des contraintes sur la cardinalité des relations.

Au niveau ontologique, les ontologies peuvent se distinguer sur le plan de la conceptualisation (Visser et al., 1997) , ce qui implique que la définition des classes, ou concepts, peut varier. Par exemple, une classe « maison » possède des attributs *adresse* et *nombre d'étages* dans une ontologie et les attributs *adresse* et *année de construction* dans une autre. Des différences dans le niveau de granularité atteint constituent aussi des hétérogénéités sur le plan de la conceptualisation. Les autres hétérogénéités au niveau

ontologique comprennent les différents paradigmes dans la façon de concevoir une même réalité, différentes descriptions de concepts, présence de synonymes, d'homonymes ou les variations lors de l'encodage.

Toutefois, l'hétérogénéité des langages, des conceptualisations et des représentations n'est pas la seule problématique liée au problème de l'interopérabilité, ce problème étant également lié à l'évolution de l'ontologie.

Toute conceptualisation ou représentation de la réalité est sujette à changement, par conséquent, l'évolution n'est pas une problématique propre à la structure multidimensionnelle mais trouve son analogue dans l'évolution des ontologies. La réalité qu'une ontologie vise à décrire et le point de vue des utilisateurs d'une ontologie sont des aspects qui sont constamment modifiés.

Les ontologies doivent être modifiées lorsqu'elles ne satisfont plus aux besoins d'une communauté scientifique en tant que référentiel sémantique, notamment :

- Pour intégrer de nouveaux concepts et de nouvelles propriétés ;
- Pour introduire un nouveau système de classification ;
- Lors de la modification des contraintes ;
- Lors de la modification du domaine de définition ou du domaine de valeur des propriétés.

L'évolution des ontologies génère des impacts négatifs sur la préservation de l'interopérabilité entre les ontologies locales (Maedche, Motik et Stojanovic, 2003), puisque les liens qui pouvaient unir les différentes ontologies ne sont pas automatiquement maintenus.

L'évolution de l'ontologie peut être définie comme la modification appropriée de l'ontologie et la propagation consistante des changements dans les artefacts dépendants, c'est-à-dire dans les objets référencés, les ontologies dépendantes, et les applications logicielles utilisant l'ontologie (Maedche et al, 2003). Cette définition décrit l'évolution de

l'ontologie uniquement sur le plan du processus en soi. Une définition plus adéquate considère que l'évolution est un processus adaptatif, c'est-à-dire qu'elle consiste en la capacité à gérer les changements de l'ontologie et leurs impacts en créant et en maintenant plusieurs versions de l'ontologie (Klein, Noy, 2003).

Nous mentionnons quelques méthodologies suggérées dans la littérature pour gérer l'évolution des ontologies.

La méthodologie proposée par AIFB (Maedche et al, 2003) suggère cinq phases qui doivent être mises en oeuvre pour gérer l'évolution de l'ontologie. La première phase a pour objet de représenter les changements générés par l'évolution. La représentation des changements s'effectue au moyen de types d'opérations qui sont analogues aux classes de modifications mentionnées dans la taxonomie de l'évolution de la structure multidimensionnelle. Deux types de changements sont distingués : les changements élémentaires et les changements complexes. Les premiers correspondent à des opérations qui ne peuvent être décomposées en opérations plus simples ; ils comprennent les ajouts suppressions et modifications des entités ontologiques (classe d'entités, attributs, relations entre les classes d'entités). Ces opérations correspondent aux opérations simples qui peuvent être effectuées sur les membres du schéma des instances dans la gestion de l'évolution de la structure multidimensionnelle (Body et al, 2004) Le deuxième type de changement, comme son nom l'indique, est une opération composée de plusieurs changements élémentaires, à savoir le déplacement, la fusion et la séparation des entités ontologiques. Comme les changements élémentaires, ces changements complexes sont analogues respectivement aux opérations complexes de reclassification, fusion et division effectuées sur les membres du schéma des instances de la structure multidimensionnelle. L'analogie entre l'évolution du schéma de la structure multidimensionnelle et l'évolution de l'ontologie est illustrée dans le tableau 2.1. Les entités ontologiques peuvent être des concepts ou des instances de ces concepts. Les concepts sont des objets généraux qui peuvent être instanciés ; ils correspondent à des niveaux dans la structure multidimensionnelle.

Évolution de la structure multidimensionnelle	Évolution des ontologies
Au niveau du schéma des instances : • Ajout ou suppression de membres ; • Modification des membres ; • Reclassification de membres • Fusion de membres ; • Division de membres. **Au niveau du schéma de la dimension** : • Ajout ou suppression de niveau ; • Changement de position des niveaux	• Ajout ou suppression d'entités ; • Modification d'entités ; • Déplacement des entités dans le graphe; • Fusion d'entités; • Séparation d'entités.

Tableau 3.1 : Taxonomie des évolutions dans la structure multidimensionnelle et les ontologies

Les modifications de l'ontologie qui sont mises en œuvre lors de la première phase peuvent engendrer des inconsistances. Une ontologie est consistante si les entités ontologiques et les valeurs respectent les contraintes imposées par le modèle ontologique. La seconde phase de la méthodologie consiste, par conséquent, à s'assurer que la consistance des ontologies est conservée en ajustant ou en ajoutant les contraintes nécessaires. La phase de l'implémentation consiste à exécuter les changements une fois qu'ils ont été validés par un utilisateur. La phase suivante est la propagation des changements appliqués à l'ontologie lors des phases précédentes aux ontologies qui sont dépendantes de celle-ci. La dernière phase consiste à évaluer les changements et à les corriger si cela s'avère nécessaire.

Cette méthodologie demeure toutefois limitée car comme les approches de mise à jour dans les bases de données multidimensionnelles, elles ne conservent pas l'historique et se contentent de représenter l'ontologie dans sa version la plus actuelle. D'autres approches se limitent ainsi à opérationnaliser les changements dans l'ontologie (Noy, Klein, 2003).

Une seconde méthodologie est identifiée au principe du versioning (Klein, Noy, 2003). Ce principe énonce que la gestion de l'évolution de l'ontologie exige de pouvoir conserver et accéder aux différentes versions de l'ontologie qui ont été générées au cours de l'évolution de celle-ci, ainsi que de permettre de spécifier par un modèle les relations qui rendent compte des changements effectués entre les versions. Le modèle doit pouvoir expliciter les relations sémantiques qui existent entre les entités ontologiques de deux versions et ainsi permettre de détecter les changements qui se sont produits lors de l'évolution, les informations sur les modifications, telles que leur durée de validité, devant être maintenues

sous forme de métadonnées. La méthodologie proposée ne suggère pas cependant comment les versions de l'ontologie peuvent être crées. Une approche complète devrait donc tenir compte du cadre défini par les deux approches précédentes.

Une approche qui englobe ces deux dernières a été proposée et intègre un formalisme de représentations de l'évolution des ontologies sous forme de graphes orientés (Eder, Koncilia, 2004). Elle se fonde sur les concepts et les techniques développées pour les bases de données temporelles, l'évolution de schéma et le versioning des bases de données, c'est-à-dire qu'un graphe regroupe toutes les versions en intégrant la notion de temps valide dans la définition des classes et des relations qui forment le graphe. La création de version est produite au moyen d'opérateurs qui ajoutent, suppriment ou mettent à jour (modifient) les relations et les classes. Cette approche, bien que plus complètes que les précédentes, demeure limitée car elle ne permet pas de découvrir les liens entre les entités ontologiques de différentes versions préexistantes.

Le principe du versioning soulève, dans un contexte différent de celui de l'hétérogénéité du langage, de la conceptualisation et des représentations ontologiques, la problématique de la diminution de l'interopérabilité des ontologies qui se manifeste également à un autre niveau, à savoir entre les versions d'une même ontologie. Bien que les versions de l'ontologie préservent l'historique de celle-ci, elles ne sont pas, a priori, complètement interopérables. C'est pourquoi l'approche de versioning implique qu'un modèle doit pouvoir représenter les relations entre les entités de deux versions de l'ontologie. Cet important élément du cadre méthodologique amène à considérer les approches de mapping entre les ontologies en tant que méthodologie pour établir les liens entre les versions.

3.2.5 Mapping des ontologies

Malgré l'hétérogénéité des ontologies, celles-ci doivent être liées entre elles pour rendre possibles les tâches d'intégration, de partage d'information, de recherche d'information à partir de plusieurs sources, etc. Le besoin de combiner des ontologies développées de façon indépendante et comportant des hétérogénéités a soulevé des problématiques sur le plan du langage des ontologies, de la conceptualisation et de la spécification (Visser et al, 1997), ainsi que des problématiques liées à l'évolution et donc au versioning des ontologies

(Fernandez et al, 1997; Klein, 2001 ; Charlet et al, 2003 ; Eder, Koncilia, 2004 ; Sure et al, 2004).

Plusieurs types de solution ont été proposés pour la résolution de l'hétérogénéité entre les ontologies. D'une part, les **approches globales** mettent l'emphase sur la construction d'ontologies volumineuses et standardisées, dans le but de produire des bases de connaissances réutilisables et ayant une durée de vie appréciable. C'est le cas de l'intégration et de la fusion des ontologies (Noy, Musen, 2000 ; Chalupsky, 2000 ; Mc Guinness et al, 2000) Certains auteurs argumentent que la problématique de l'hétérogénéité ne peut être résolue par les méthodes de fusion des ontologies, qui sont trop coûteuses et prennent rapidement de l'ampleur (Mitra, Wiederhold, 2002). À l'opposé, on trouve donc un autre type de solution, celle des **ontologies multiples**, visant à conserver la structure des différentes ontologies en cherchant plutôt à établir des relations entre celles-ci. Entre ces solutions se trouve le compromis de **l'approche hybride**, où la sémantique de chaque source est conservée et décrite par sa propre ontologie, mais où toutes les ontologies sont construites à partir d'un vocabulaire commun, les primitives. Celles-ci représentent les termes ou concepts de base qui permettent de construire les concepts complexes. Cependant, dans cette approche, les ontologies existantes ne peuvent être utilisées directement mais doivent être développées à nouveau pour être conséquentes avec le vocabulaire commun. Ces trois types d'approches sont illustrés par la figure 2.1.

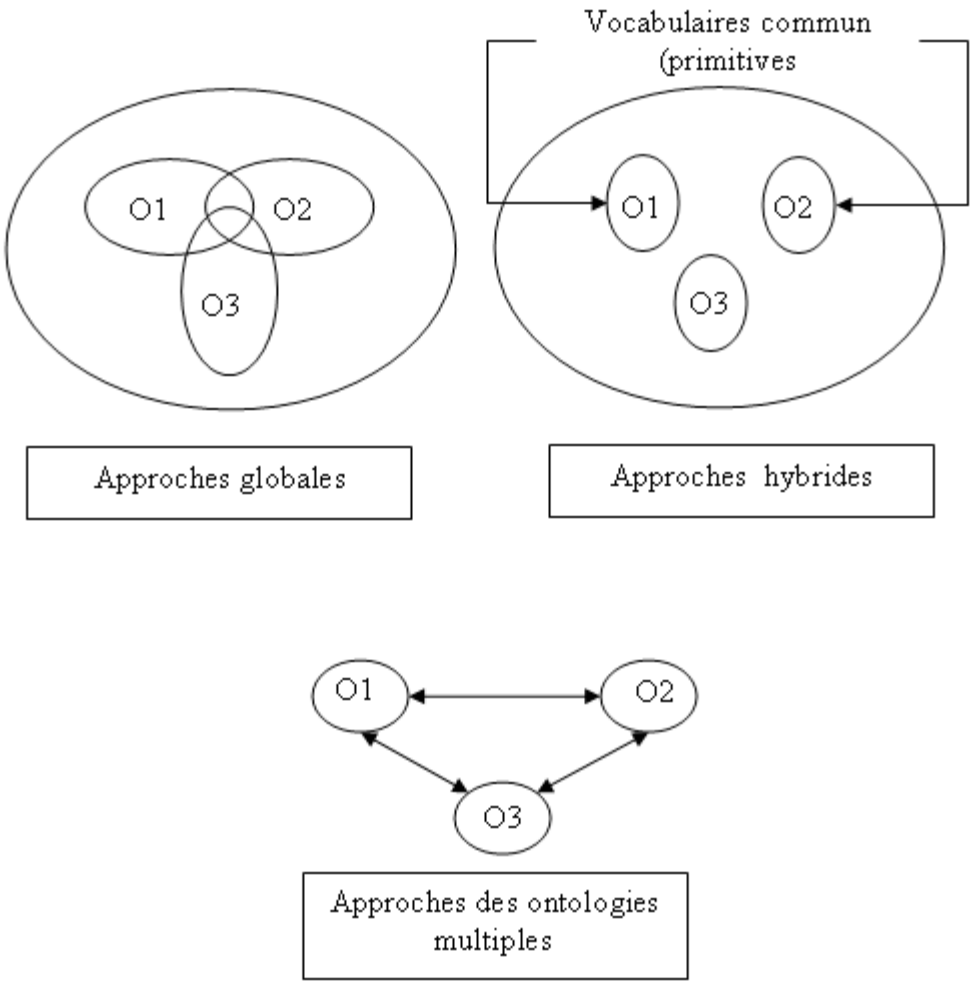

Figure 3.2 : Les trois types d'approches pour la gestion de l'hétérogénéité entre les ontologies

La fusion, ou intégration, d'ontologies consiste à combiner une ou plusieurs ontologies qui utilisent des vocabulaires différents mais dont une partie ou la totalité des contenus sont communs (McGuinness et al., 2000). Des cadres de travail sont élaborés pour créer des conventions visant à rendre homogène la conceptualisation et la spécification des ontologies (Brickley et al., 1999) ; on vise alors à produire une ontologie globale et générale qui peut constituer une base pour la conception d'ontologies spécifiques aux différentes applications (Niles et Pease, 2001; Gangemi et al., 2003). Cette ontologie de référence facilite alors la recherche de correspondances entre les ontologies qui en sont dérivées. Ce mécanisme implique, d'une part, l'identification des zones communes des ontologies par l'établissement de liens entre les entités similaires sur le plan sémantique ;

celles-ci doivent ensuite référer à la même entité dans l'ontologie résultante. La seconde phase consiste à identifier les entités qui peuvent être liées par le mécanisme d'alignement. Ce mécanisme consiste à modifier les ontologies pour les rendre conciliables (Compatangelo et Meisel, 2002).

L'approche des ontologies multiples a pour objet de développer une fonction de mapping qui puisse établir les liens entre les entités ontologiques. La recherche de fonction de mapping, qui est l'approche adoptée dans notre cadre de travail, met en relation, au moyen d'une métrique, les concepts et relations similaires appartenant à deux ontologies, et pourra être utilisée pour mettre en relation les membres des schémas d'instances des dimensions appartenant à des bases de données multidimensionnelles différentes. L'approche des ontologies multiples diffère de l'alignement des ontologies dans le fait qu'elle ne modifie pas les concepts pour produire un ajustement. Les approches de recherche de mapping diffèrent également selon le paradigme adopté pour découvrir les relations. Les méthodes heuristiques emploient les caractéristiques des ontologies pour établir les mapping, telles que la structure et les caractéristiques lexicales (Noy, Musen, 2003). Certaines s'appuient sur la recherche de structures invariantes (isomorphismes) pour détecter les mapping (Bench-Capon et Malcolm, 1999); par exemple, la théorie du flux d'information (Barwise et Seligman, 1997) a inspiré des travaux dans le domaine des mapping entre les ontologies où la recherche des mapping est automatisée et est formalisé en tant qu'infomorphismes logiques (Kalfoglou et Shorlemmer, 2002). Par ailleurs, plusieurs de ces approches sont orientées vers les besoins spécifiques de l'utilisateur et exploitent les entrées de l'utilisateur pour ajuster le mapping (Lacher et Groh, 2001; Noy et Musen, 2003). Ainsi, de façon plus poussée, le système GLUE (Doan et al., 2002) intègre un système d'apprentissage-machine exploitant une méthode probabiliste pour modéliser le comportement des utilisateurs et crée un système de mapping basé sur plusieurs mesures de similarité.

3. 3 Notion de similarité sémantique

Afin de pouvoir mettre en relation des ensembles de concepts, il est nécessaire de disposer d'une mesure qui permet d'en évaluer, de manière quantitative ou qualitative, la similitude et la dissimilitude. La notion de similarité sémantique est à la base de la définition de cette

mesure. Toutefois, il importe de distinguer entre la notion de relation sémantique et la notion de similarité. Deux concepts sont similaires s'ils atteignent un certain niveau de ressemblance ; des concepts dissimilaires peuvent également être liés sémantiquement par des relations lexicales : métonymie, antonymie, spécialisation, etc. La fonction de similarité sémantique est définie par les éléments suivants (Bisson, 1998):

- une représentation des concepts et le langage utilisé ;
- une représentation de la de similarité ;
- le contexte ou l'ensemble des connaissances sur l'univers de discours étudié;
- une fonction binaire de similarité.

De manière générale, la similarité entre deux entités est évaluée à partir du degré de chevauchement des entités en fonction de leurs propriétés. Pour des entités identiques, la similarité prend une valeur maximale.

3.3.1 Propriété de la distance de similarité

L'élaboration de modèles pour la mesure de similarité sémantique doit se faire sur la base de propriétés qui sont attribuées à la notion de similarité. Quelques modèles considèrent que la similarité doit être une distance (métrique).

Une distance métrique doit respecter les critères suivants:

1) La distance entre un élément et lui-même est nulle: $\delta(c_i, c_i) = 0;$ (3.1)

2) Propriétés de l'inégalité triangulaire: $\delta(c_i, c_j) + \delta(c_j, c_k) \geq \delta(c_i, c_k)$; (3.2)

3) Symétrie: $\delta(c_i, c_j) = \delta(c_j, c_i)$; (3.3)

Cependant les propriétés énoncées pour la similarité sémantique relèvent plutôt du domaine de la psychologie cognitive ; en général on reconnaît que la première propriété doit être respectée car la distance entre un objet et lui-même doit toujours être la plus petite distance possible ; quelques modèles donnent toutefois une mesure non nulle de la distance entre un concept et lui-même, où la mesure dépend des caractéristiques de l'ontologie (Resnick,

1999). La propriété de symétrie n'est pas toujours reconnue comme une caractéristique de la similarité puisqu'il est considéré par certains qu'un objet particulier est plus similaire à son prototype, plus général, que ce prototype est similaire à cet objet particulier, (Tversky, 1977) parce qu'en fait l'objet particulier possède toute les caractéristiques de l'objet prototype mais que l'inverse n'est pas vérifié.

3.3.2 Modèles de similarité sémantique

Les approches de mapping heuristiques s'appuient sur des mesures de similarité sémantique pour établir les relations entre les concepts ; il existe dans la littérature plusieurs modèles de mesure de la similarité sémantique qui peuvent être regroupés selon la représentation des concepts qu'ils utilisent. D'une part, les modèles s'appuient sur une représentation qualitative ou quantitative des concepts. Dans une représentation qualitative, les concepts possèdent des propriétés descriptives alors que dans les modèles quantitatifs les propriétés sont numériques ou quantifiables. Les modèles de similarité sémantique emploient une mesure de similarité, une distance de similarité ou encore une mesure qui combine ces deux dernières.

Les mesures de similarité prennent pour objets les concepts qui composent une ontologie. Elles peuvent également être classifiées selon la représentation sur laquelle elles s'appuient: arbre de la taxonomie de l'ontologie, représentations dans un espace vectoriel ou propriétés des concepts. Certaines de ces mesures emploient aussi la notion de transformation en évaluant l'importance des modifications nécessaires devant être effectuées pour obtenir un premier concept à partir du second.

3.3.2.1 Modèles basés sur les graphes

Les modèles basés sur la structure arborescente de la taxonomie s'appuient sur la théorie *spreading activation theory* qui pose l'hypothèse que la hiérarchie des concepts est organisée selon les lignes de similarité sémantique. Par conséquent, des concepts de l'ontologie, liés par des arcs, sont similaires si la distance qui les sépare dans le graphe est faible, la distance dans un graphe étant donnée par le plus court chemin à parcourir le long des arcs pour joindre un concept à partir d'un autre. Ce type de mesure a été introduit par

Rada (Rada et al, 1989) qui propose comme mesure de similarité entre les concepts c_1 et c_2 :

$$\mathrm{sim}(c_1, c_2) = \frac{1}{1 + \mathrm{dist}(c_1, c_2)} \tag{3.4}$$

La distance entre deux noeuds n'est pas nécessairement uniforme (Jiang et Conrath, 1998) ; d'autres mesures intégrant ces potentialités ont donc été introduites et tiennent compte de la densité locale des nœuds, de la profondeur des concepts dans le graphe, de la profondeur totale du graphe et de la force des relations. La mesure de Resnik (Resnik, 1995) incorpore la profondeur maximale (max) du graphe :

$$\mathrm{sim}(c_1, c_2) = 2\max - \mathrm{dist}(c_1, c_2) \tag{3.5}$$

Cette mesure a été élaborée en tenant compte du fait que la distance conceptuelle doit se comporter comme une métrique, c'est-à-dire respecter les propriétés de symétrie, transitivité et d'inégalité triangulaire. Ces propriétés ne font toutefois pas l'unanimité dans la caractérisation de la mesure de similarité sémantique.

La mesure de Wu et Palmer (Wu et Palmer, 1994) donne la similarité entre concepts appartenant à un domaine conceptuel en tenant compte de la profondeur du plus petit généralisateur commun prof(c) :

$$\mathrm{sim}(c_1, c_2) = \frac{2\mathrm{prof}(c)}{\mathrm{prof}(c_1) + \mathrm{prof}(c_2)} \tag{3.6}$$

La mesure de Hirst et St-Onge (Hirst et St-Onge, 1998) élaborent une mesure de similarité entre les mots en pondérant la distance entre les concepts selon le type de relation (relations de généralisation, antonymes, synonymes, etc.). La similarité est élevée si le chemin entre les concepts est court et que la direction des arcs reliant les concepts varie peu :

$$\mathrm{sim}(c_1, c_2) = C - L - kd \tag{3.7}$$

avec L la longueur du chemin et d le nombre de changement de direction, C et k constants. Li (Li et al, 2003) suggère une mesure de similarité non linéaire pour tenir compte du fait

que la quantité d'information est infinie mais que l'intervalle de similarité, compris entre 0 et 1, est fini. La transformation d'un domaine à l'autre doit donc être non linéaire :

$$\mathrm{sim}(c_1, c_2) = e^{-\alpha L} \frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \tag{3.8}$$

$\beta>0$ et $\alpha\geq 0$ sont des paramètres qui pondèrent la profondeur du premier généralisateur commun H et L le plus court chemin entre c_1 et c_2.

Les mesures présentées précédemment permettent d'évaluer la similarité entre deux concepts (nœuds) situés dans un graphe. D'autres approches emploient des mesures qui ont été conçues pour évaluer la similarité entre deux graphes afin de déterminer si les graphes sont identiques (isomorphes), si un des graphes englobe le second ou à quel degré des graphes sont dissimilaires. La mesure de similarité élaborée par Champin et Solnon (Champin et Solnon, 2003), qui se base sur le modèle ratio de Tversky (Tversky, 1977) permet de résoudre ces trois situations en optimisant la valeur de la similarité pour obtenir le mapping optimal entre les nœuds et entre les arcs des graphes ; de plus, cette approche tient compte de la possibilité d'existence de *matching* multiples. D'autres approches conçoivent une mesure de similarité qui prend en paramètre les modifications qui rendront deux graphes isomorphes (Bunke, 1999) ou le nombre de sous-graphes entièrement connectés (Coulon, 1995).

3.3.2.2 Modèles basés sur le contenu informatif

Ces modèles quantifient le fait que plus le contenu informatif partagé par deux concepts est élevé, plus ces concepts sont similaires ; ils s'appuient également sur la structure taxonomique de l'ontologie. Le contenu informatif d'un concept *c* est fonction de la probabilité d'occurrence (probabilité de présence) des composantes de ce concept, laquelle dépend de la fréquence de ses composantes dans l'ensemble de l'ontologie. Cette fréquence est définie par:

$$\mathit{fréq}(c) = \sum_{n \in C} n \tag{3.9}$$

Avec C l'ensemble des termes composant le concept c. La probabilité est une fonction $p : c \rightarrow [0,1]$ telle que :

$$P(c) = \frac{fréq(c)}{N} \tag{3.10}$$

Où N est le nombre d'instances de tous les concepts dans l'ontologie. Cette fonction indique que plus la fréquence des composantes du concept *c* est importante dans l'ontologie, plus la probabilité d'occurrence de c est élevée. Finalement, le contenu informatif du concept c est une fonction logarithmique de la probabilité P(c):

$$\text{contenu en information} = -\log(\mathrm{P(c)}) \tag{3.11}$$

La fonction logarithmique est négative et décroît quand son argument est plus petit ou égal à 1. Par conséquent, plus la probabilité P(c) d'un concept est faible, plus son contenu informatif est élevé et inversement plus la probabilité est forte, plus son contenu informatif est faible. L'équation du contenu en information exprime que plus le concept est général (à un niveau élevé dans la taxonomie), plus son contenu en information est faible car sa probabilité d'occurrence est forte et inversement plus le concept est spécialisé plus son contenu en information est important car la probabilité d'occurrence est plus faible. À partir de cette définition et du principe de l'approche du contenu informatif, la similarité entre deux concepts c1 et c2 sera donnée par le contenu en information qu'ils partagent, c'est-à-dire par celui du premier généralisateur commun c (Resnik, 1999):

$$\mathrm{sim}(c_1, c_2) = \max_{c \in S(c_1, c_2)} (\log \mathrm{P(c)}) \tag{3.12}$$

où S est l'ensemble des généralisateurs commun de c_1 et c_2. Le premier généralisateur commun de c_1 et c_2 est le parent commun de c_1 et c_2 dont le contenu informatif est le plus élevé parmi tous les parents de c_1 et c_2. Cette mesure donne pour valeurs maximale - ln (1/N) = ln (N) (pour des concepts parfaitement similaires) et une valeur qui tend vers 0 pour des termes qui n'ont aucun contenu similaire. Notons que la mesure de la similarité d'un concept avec lui-même ne donne pas la valeur de similarité maximale possible dans le corpus, mais qu'elle dépend de la profondeur du concept dans la taxonomie. Deux concepts différents peuvent donc, selon cette mesure, sembler plus similaires qu'un concept avec lui-même, ce qui fait que l'usage de cette mesure doit être accompagné de précautions particulières dans son interprétation. Cette mesure a toutefois pour avantage de fournir de

l'information sur la position d'un concept dans la taxonomie et de pouvoir indiquer la taille de l'ontologie, à savoir N le nombre d'instances des concepts.

Une autre mesure tient compte non seulement du contenu informatif commun entre deux concepts mais également de leur propre contenu informatif (Lin, 1993):

$$sim(c_1, c_2) = \frac{2\ln(P(c))}{\ln(P(c1)) + \ln(P(c2))} \tag{3.13}$$

Cette normalisation permet de réduire le domaine de valeur de la similarité à l'intervalle [0,1], une valeur de 1 indiquant des concepts parfaitement similaires. Cette mesure permet de faire des distinctions plus fines des valeurs de similarité que la mesure précédente.

La distance sémantique suivante conserve les avantages des deux mesures précédentes (Jiang et al, 1998):

$$dist(c_1, c_2) = -2\ln P(c) - (\ln P(c_1) + \ln P(c_2)) \tag{3.14}$$

C'est-à-dire qu'elle peut indiquer la taille de l'ontologie par sa valeur maximale et qu'elle prend aussi en compte le contenu informatif de chaque concept.

Selon une autre approche, le contenu en information d'un concept est une fonction du nombre d'hyponymes de ce concept, hypo(c), et du nombre de concepts dans la taxonomie $\max_{wn}$ (Seco et al, 2004):

$$\mathrm{ic(c)} = 1 - \frac{\log(\mathrm{hypo(c)} + 1)}{\log(\max_{\mathrm{wn}})} \tag{3.15}$$

3.3.2.3 Modèles basés sur une représentation vectorielle

Les modèles de représentation vectorielle utilisent une analogie où la proximité sémantique entre concept est représentée par la proximité spatiale dans un espace vectoriel. Ce modèle est surtout utilisé en recherche d'information pour représenter des documents. Les n dimensions de l'espace vectoriel correspondent alors à l'ensemble des n termes (mots ou groupes de mots) qui constituent l'ensemble du corpus, et les documents sont représentés

par des vecteurs V_D dans cet espace, où chaque composante v_i correspond à la fréquence d'un terme dans le document. À partir de cette représentation, plusieurs mesures de similarité ont été élaborées, telles que le coefficient de Jacquard qui intègre les éléments communs entre les objets au numérateur et les éléments différents entre les objets au dénominateur:

$$sim(D_1, D_2) = \frac{\sum_{i=1}^{n} v_{1i} v_{2i}}{\sum_{i=1}^{n} v_{1i} v_{1i} + \sum_{i=1}^{n} v_{2i} v_{2i} - \sum_{i=1}^{n} v_{1i} v_{2i}} \tag{3.16}$$

Un autre modèle est celui du cosinus de Salton (Salton, Buckley, 1990):

$$sim(D_1, D_2) = \frac{\sum_{i=1}^{n} v_{1i} v_{2i}}{\sqrt{\sum_{i=1}^{n} v_{1i}^2 \sum_{i=1}^{n} v_{2i}^2}} \tag{3.17}$$

La similarité de Salton calcule le cosinus de l'angle entre les vecteurs en se basant sur leur produit scalaire. Quand les documents sont identiques, l'angle entre les vecteurs est nul et le cosinus vaut 1 et à l'opposé des documents entièrement différents sont représentés par des vecteurs orthogonaux donc leur similarité est nulle. Ces modèles se basent sur une représentation des concepts, ou documents, par des éléments discrets (les composantes des vecteurs) Le modèle de similarité peut être étendu au cas où les propriétés des concepts sont quantitatives et continues ; dans ce cas la fréquence de chaque propriété devient aussi une fonction continue, la fonction de densité ρ, et la sommation pour le calcul des éléments communs est transformée en intégrale :

$$sim(c_1, c_2) = \int \rho_1(r) \rho_2(r) dr \tag{3.18}$$

Le domaine de valeur de la similarité peut être ramené à l'intervalle [0,1] en normalisant avec l'équation du cosinus de Salton (Carbo, 1998):

$$sim(c_1, c_2) = \frac{s(c_1, c_2)}{\sqrt{s(c_1, c_1) s(c_2, c_2)}} \tag{3.19}$$

La mesure de similarité employée par les modèles vectoriels est transitive et symétrique.

3.3.2.4 Modèles basés sur les propriétés

Les modèles qui s'appuient sur les propriétés des concepts s'appuient sur une base qualitative ; ils sont fondés sur la comparaison du nombre de propriétés communes par rapport au nombre total de propriétés. Établissant leur base sur la théorie ensembliste, ils produisent une mesure de similarité qui n'est pas nécessairement une métrique car les propriétés de symétrie et de transitivité ne sont pas toujours respectées. Le modèle ratio de Tversky (Tversky, 1977) prend en compte le nombre de propriétés communes et les différences entre les deux concepts comme dans la mesure du coefficient de Jaccard:

$$S(c_1, c_2) = \frac{f(C_1 \cap C_2)}{f(C_1 \cap C_2) + \alpha f(C_1 - C_2) + \beta f(C_2 - C_1)} \tag{3.20}$$

Où C_1 est l'ensemble des propriétés de c_1 et C_2 est l'ensemble des propriétés de c_2. Avec une fonction f qui est croissante monotone et $\alpha \geq 0, \beta \geq 0$ sont des paramètres qui pondèrent les différences. La différence C_1 - C_2 correspond aux propriétés possédées par C_1 mais absentes de C_2 et $C_2 - C_1$ correspond aux propriétés possédées par C_2 mais absentes de C_1. Ce modèle permet de donner une mesure asymétrique en attribuant un poids différent pour les différences car il n'est pas toujours souhaitable que la relation de généralisation, par exemple, indique une même valeur de similarité que la relation de spécialisation.

3.3.2.5 Modèles hybrides

Plusieurs approches combinent les propriétés des modèles précédents pour produire une mesure de similarité qui considère plusieurs aspects de la représentation des concepts. La mesure de similarité de Knappe (Knappe, 2003) prend en compte l'ensemble des concepts qui sont liés aux deux concepts comparés dans le graphe de l'ontologie. Cette mesure emploie les attributs des concepts ainsi que leur profondeur dans le graphe. Cette approche est intéressante car elle considère que les concepts peuvent être simples (atomiques) ou composé (formés de plusieurs concepts). Soit $u(c_1)$ l'ensemble des concepts atteignables par c_1 et plus généraux que c_1 et soit $u(c_2)$ l'ensembles des concepts atteignables par c_2 et

plus généraux que c_2. La similarité est fonction de l'intersection de ces ensembles et du paramètre $\rho \in [0,1]$:

$$\mathrm{sim}(c_1, c_2) = \rho \frac{|u(c_1) \cap u(c_2)|}{|u(c_1)|} + (1-\rho) \frac{|u(c_1) \cap u(c_2)|}{|u(c_2)|} \tag{3.21}$$

Le paramètre ρ détermine le degré d'influence de la relation de généralisation et permet de définir une fonction de similarité asymétrique. Cette mesure partage avec la mesure de Rodriguez (Rodriguez, 2000) la propriété selon laquelle la similarité est asymétrique en regard des relations de généralisation et de spécialisation, c'est-à-dire que la similarité entre une entité spécialisée et une entité générale est considérée comme plus élevée que la similarité entre la même entité générale et son entité spécialisée.

L'approche proposée par Rodriguez (2000) combine des éléments du modèle ratio de Tversky, ainsi que des facteurs qui tiennent compte du contexte, pour produire une mesure de similarité qui associe des entités spatiales appartenant à la même ontologie (modèle Matching Distance) ou à des ontologies différentes (Modèle Triple Matching Distance). Ce modèle tient compte du fait que l'évaluation de la similarité s'effectue dans un domaine de discours lequel définit un contexte et que, par conséquent, la mesure de similarité entre deux entités de classe est dépendante du contexte.

Le modèle Matching Distance s'applique à des ontologies où les classes sont organisées taxonomiquement et sont liées par des relations de généralisation (*is-a*) ou des relations d'agrégation (*part-of*). Dans le premier modèle Matching Distance, la similarité entre les entités de classes est évaluée par une somme pondérée des similarités calculées selon trois catégories de propriétés : les *attributs*, qui caractérisent intrinsèquement les entités, les *parties* qui peuvent être des propriétés intrinsèques ou des éléments ayant une relation *part-of* avec les entités, et les *fonctions* qui caractérisent le comportement et le rôle des entités.

$$S(c_1, c_2) = \omega_a S_a(c_1, c_2) + \omega_p S_p(c_1, c_2) + \omega_f S_f(c_1, c_2) \tag{3.22}$$

Les similarités sont pondérées par les poids ω_a pour la similarité des attributs, ω_p pour la similarité des parties et ω_f pour la similarité des fonctions ; ces poids ont pour rôle d'ajuster la similarité en fonction du contexte. Le contexte regroupe les entités de classe dont les propriétés présentent un intérêt pour l'utilisateur ; il peut donc être considéré comme un thème tel que ce concept a été défini dans le domaine des ontologies spatiales. En considérant les entités de classe qui constituent le contexte, la pertinence des propriétés peut être déterminée selon deux approches : soit en considérant que les propriétés pertinentes dans l'évaluation de la similarité sont celles qui portent le plus important contenu en information, ce qui correspond à l'approche de la variabilité, soit en considérant que les propriétés pertinentes sont celles qui contribuent le plus à la caractérisation du contexte, ce qui correspond à l'approche de la ressemblance. Dans le premier cas, selon le principe de l'approche de la variabilité, la pertinence des propriétés de type t (t=attributs, parties ou fonctions) est donnée par

$$P_t^v = 1 - \sum_{i=1}^{l} \frac{o_i}{n} \tag{3.23}$$

Où o_i est la fréquence d'une propriété i, $i \in t$, dans l'ensemble du contexte, l est l'ensemble des propriétés de type t dans l'ensemble du contexte et n est l'ensemble des entités de classe de l'ontologie. La somme correspond à la somme des probabilités d'occurrence dans l'ontologie des propriétés de type t présentes dans le contexte. À l'opposé, selon l'approche de la ressemblance, la pertinence est définie par

$$P_t^c = \sum_{i=1}^{l} \frac{o_i}{n} = 1 - P_t^v \tag{3.24}$$

Le poids de chaque type de propriété est donné par la pertinence, calculée selon l'une ou l'autre des approches:

$$\omega_a = \frac{P_a^{(v,c)}}{P_a^{(v,c)} + P_p^{(v,c)} + P_f^{(v,c)}}$$

$$\omega_p = \frac{P_p^{(v,c)}}{P_a^{(v,c)} + P_p^{(v,c)} + P_f^{(v,c)}}$$

$$\omega_f = \frac{P_f^{(v,c)}}{P_a^{(v,c)} + P_p^{(v,c)} + P_f^{(v,c)}} \tag{3.25}$$

Ce modèle offre donc deux possibilités pour l'évaluation du poids de chaque type de propriétés, selon la conception que l'utilisateur peut avoir de la pertinence dans un contexte donné.

La similarité selon chaque propriété est donnée par la forme suivante de la similarité du modèle ratio:

$$S(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha |C_1 - C_2| + (1-\alpha)|C_2 - C_1|} \tag{3.26}$$

Donc la fonction f correspond à la cardinalité des ensembles C_1 et C_2 et $\beta = 1-\alpha$. Le modèle Matching Distance intègre la distance entre les concepts dans le graphe de la taxonomie à l'intérieur du paramètre α, lequel est une fonction de la profondeur d_1 du concept c_1 et de la profondeur d_2 du concept c_2 par rapport à la racine de la taxonomie:

$$\alpha = \begin{cases} \dfrac{d_1}{d_1 + d_2} & \text{si } d_1 \leq d_2 \\ 1 - \dfrac{d_1}{d_1 + d_2} & \text{sinon} \end{cases} \tag{3.27}$$

Dans l'extension du modèle Matching Distance pour plusieurs ontologies (modèle Triple Matching Distance), les graphes des ontologies sont liés par une superclasse commune afin de pouvoir calculer la distance entre les concepts dans le graphe global. La similarité de voisinage des concepts dans le graphe et la similarité lexicale (similarité entre les noms des concepts) sont ajoutées à la somme pondérée du modèle Matching Distance. Le voisinage d'un concept est l'ensemble des entités de classe qui se trouvent à un rayon r (défini selon les besoins) du concept dans le graphe. Pour un rayon assez grand, les voisinages peuvent différer de manière significative et la similarité des voisinages sera faible ; à l'opposé, un

rayon faible augmente la probabilité d'avoir un voisinage contenant plus d'éléments commun. Le choix du rayon r dépendra aussi de la taille du graphe de l'ontologie. Par exemple, pour un graphe ayant pour profondeur maximale MAX= 7, un rayon de r=3 peut sembler démesuré et diminuera inutilement la similarité ; dans une ontologie ayant pour profondeur maximale MAX=10000, un même rayon de r=3 peut être trop faible et ne sera pas représentatif du voisinage d'un concept. L'évaluation de la similarité des voisinages V_1 et V_2 de deux concepts prend une forme similaire au modèle ratio:

$$S_v(c_1,c_2) = \frac{V_1 \cap V_2}{V_1 \cap V_2 + \alpha(c_1,c_2) * \delta(c_1, V_1 \cap V_2) + (1-\alpha(c_1,c_2)) * \delta(c_2, V_1 \cap V_2)} \quad (3.28)$$

L'intersection entre les voisinages est approximée par la similarité maximale des entités de classe entre les voisinages:

$$V_1 \cap V_2 = \left[\sum_{i<n} \max_{j<m} S(V_{1i}, V_{2j}) \right] - \varphi S(c_1,c_2) \quad (3.29)$$

où V_{1i} est un élément de V_1, de cardinalité n, et V_{2j} est un élément de V_2, de cardinalité m, et

$$\varphi = \begin{cases} 1 \text{ si } S(c_1,c_2) = \max_{j \le m} S(c_1, V_{2j}) \\ 0 \quad \text{sinon} \end{cases} \quad (3.30)$$

Afin de respecter la propriété selon laquelle l'intersection des deux ensembles est toujours plus petite ou égale au plus petit des deux ensembles. La différence est donnée en fonction de l'intersection:

$$\delta(c_1, V_1 \cap V_2) = \begin{cases} n - V_1 \cap V_2 & \text{si } n > V_1 \cap V_2 \\ 0 & \text{sinon} \end{cases} \quad (3.31)$$

Le modèle Triple Matching Distance englobe donc plusieurs aspects (contexte, propriétés des concepts, distance dans les graphes, similarité des voisinages dans les graphes et similarité lexicale) qui peuvent influencer la similarité, au lieu de se concentrer sur un seul aspect des concepts qui peut être plus ou moins représentatif. Une mesure de similarité qui

ne tient compte que de la position des concepts dans un graphe ou de leur contenu en information, qui est lui-même défini de façon probabiliste, apparaît limitée dans l'optique d'une approche globale comme la nôtre où les concepts sont définis sur plusieurs plans (propriétés, position dans la hiérarchie). Le modèle Triple Matching Distance offre aussi la possibilité de comparer les voisinages des concepts dans les graphes ; ceci n'est pas pris en compte dans la plupart des modèles où la mesure de similarité est conçue pour quantifier la similarité sémantique entre des concepts qui appartiennent à la même ontologie et donc au même graphe, sauf bien sûr dans les modèles spécialement dédiés à la comparaison des graphes. Par contre, ces derniers ne prennent pas nécessairement en compte les autres aspects (propriétés et types de propriétés, contexte, par exemple) et demeurent incomplets. La similarité entre les graphes et les parties de graphes est importante dans notre contexte puisque nous désirons évaluer la similarité entre des éléments qui appartiennent à des graphes différents, à savoir des schémas de dimension dont la structure a évoluée d'une époque à l'autre. Par contre, le niveau de définition du modèle Matching Distance peut être inapproprié dans le cas où les attributs, parties et fonctions peuvent prendre des valeurs différentes et s'exprimer dans des formats complexes comme des textes ou des intervalles numériques. Par exemple, deux concepts *maison* peuvent posséder les attributs communs *surface*, *durée de l'hypothèque*, *description des pièces*, qui peuvent prendre plusieurs valeurs, par conséquent il faut aussi tenir compte de la valeur de ces attributs dans l'évaluation de la similarité, ce qui n'est pas pris en compte dans le modèle Matching Distance.

Les modèles combinant plusieurs aspects des concepts (propriétés, position dans la hiérarchie, contexte) ont pour avantage d'être complets et de prendre en compte le maximum d'information contenue dans l'ontologie ; par contre, bien qu'elles soient englobantes, elles auraient avantage à s'appliquer à un plus fin niveau qu'entre les concepts, soit entre les propriétés des concepts.

D'autres modèles de similarité sémantique se situent dans la catégorie des modèles qualitatifs, dont nous donnons un exemple ci-dessous.

3.3.2.6 Proximité géosémantique

La proximité géosémantique est une mesure qualitative de la similitude entre des concepts qui possèdent une représentation géométrique (Brodeur, 2004); elle est utilisée dans le cadre d'une approche qui vise à établir l'interopérabilité entre des bases de données géospatiales. Le modèle de mesure de proximité établit un parallèle entre la représentation des concepts et les objets spatiaux ; les concepts possèdent des propriétés de deux types, à savoir des propriétés intrinsèques (attributs descriptifs, nom, géométrie, temporalité) et des propriétés extrinsèques (les relations avec les autres concepts ainsi que les fonctions du concepts). Les propriétés intrinsèques sont associées à l'intérieur du concept alors que les propriétés extrinsèques sont associées à la frontière du concept. La proximité géosémantique évalue l'intersection des intérieurs et des frontières des deux concepts et produit donc seize prédicats pour la proximité géosémantique en se basant sur les relations topologiques entre les entités spatiales. La proximité permet par conséquent de déterminer si un concept est contenu dans l'autre, si des concepts se chevauchent, s'ils sont égaux, etc. Elle constitue un bon outil pour lier des concepts de deux ontologies différentes, en produisant des résultats qualitatifs qui, une fois ordonnés, permettront de répondre à des requêtes. L'avantage de cette méthode réside dans son aspect qualitatif qui donne une définition plus large du concept. Par conséquent, la quantification de ce modèle, si elle est combinée avec l'information qualitative qu'il fournit, peut contribuer à le rendre plus précis par rapport aux modèles exclusivement quantitatifs cités plus haut.

3. 4 Conclusion

Après avoir revisité les concepts de base de l'ontologie ainsi que le problème de leur évolution, il a été possible d'établir un parallèle entre l'évolution de l'ontologie et celle de la structure multidimensionnelle. En particulier, les approches pour la gestion de l'évolution de l'ontologie, qui consistent à modifier l'ontologie en fonction des changements ou à utiliser le principe du versioning, sont analogues aux méthodes mises en œuvre pour résoudre le problème de l'évolution du modèle multidimensionnel que sont les mises à jour et les approches à versions. Dans les deux cas, les approches de versioning sont appropriées car elles permettent d'atteindre le but visé qui est de conserver l'historique; cependant elles engendrent la nécessité d'établir les liens entre les versions,

soit pour maintenir l'interopérabilité entre les versions dans le cas des ontologies, ou pour permettre de poser des requêtes évolutives (portant sur plusieurs versions) dans le cas de la structure multidimensionnelle. C'est dans ce but que les méthodes de mapping ont été explorées, c'est-à-dire pour pouvoir réaliser automatiquement les *mappings* entre les versions ; plus particulièrement, nous nous sommes intéressés au cas où le mapping s'effectuait au moyen d'une mesure de similarité entre les concepts, cette mesure de similarité pouvant se baser sur divers aspects de la représentation des concepts, soit sur les graphes, le contenu en information, la représentation vectorielle, les propriétés des concepts, ou encore combiner plusieurs de ces aspects, à savoir les modèles hybrides. Nous avons donc envisagé qu'une mesure de similarité puisse être élaborée pour évaluer la similarité entre les membres (instances) des schémas d'instances des dimensions des différents cubes afin de rétablir les liens sémantiques entre ceux-ci.

Puisque les ontologies ont été conçues afin de représenter la sémantique des données, les approches qui sont basées sur leur utilisation apparaissent pertinentes pour résoudre le problème de l'évolution de la sémantique dans la structure multidimensionnelle. Plus particulièrement, nous proposons de nous inspirer de la méthodologie des mapping qui utilise une mesure de similarité entre les ontologies pour rétablir les liens entre les membres du schéma des instances de la structure multidimensionnelle. Afin de prendre en compte le maximum d'information sur ces membres dans l'évaluation de la similarité, nous choisissons une mesure de similarité hybride, le modèle Matching Distance, que nous adaptons afin de pouvoir évaluer la similarité à un niveau plus fin qui prend en compte la présence d'attributs complexes tels que des textes et des intervalles de valeurs. L'approche proposée pour le mapping sémantique entre les versions de la structure multidimensionnelle est présentée dans le chapitre suivant.

Chapitre 4 : Similarité sémantique et redéfinition du modèle de similarité

4. 1 Introduction

Dans l'optique du développement d'une approche géosémantique intégrée, nous avons développé des méthodes qui permettent de rétablir les liens sémantiques et géométriques entre différents cubes de données géospatiales représentant une même réalité. Ce chapitre présente la méthode développée pour rétablir les liens sémantiques. Les ontologies étant conçues, tel que présenté dans le chapitre précédent, dans le but de représenter la sémantique des données, leur utilisation apparaît pertinente pour résoudre le problème de l'évolution de la sémantique dans la structure multidimensionnelle. L'approche de rétablissement de liens sémantiques que nous avons développée s'inspire des méthodes heuristiques de mapping entre les ontologies, c'est-à-dire des méthodes qui utilisent une fonction de similarité sémantique pour établir les relations entre les concepts de deux ontologies. Notre approche s'appuie sur un modèle de similarité sémantique, le modèle Matching Distance (Rodriguez, 2000), que nous avons redéfini et amélioré pour le rendre à la fois plus général, précis et applicable au contexte de données spatiales forestières. Notre modèle redéfini améliore la portée du modèle Matching Distance en lui permettant de mesurer non seulement la similarité entre les concepts uniquement représentés par un mot ou un ensemble de synonymes, mais également entre des concepts possédant des propriétés complexes, comme par exemple des textes ou des intervalles. L'extension que nous proposons nous a ainsi amené ainsi à concevoir une méthode de segmentation et d'indexation de textes. De plus, le modèle de similarité redéfini incorpore une nouvelle mesure conçue spécifiquement pour évaluer la similarité aux niveaux agrégés d'une hiérarchie de concepts.

Le modèle de similarité sémantique redéfini, développé sur le plan théorique pour évaluer la similarité entre des concepts, sera appliqué pour évaluer la similarité sémantique entre les membres (instances) des différents schémas d'instances des dimensions appartenant à différents cubes de données géospatiales. Les relations sémantiques établies au moyen de la

fonction de similarité permettent de créer un ensemble de matrices de correspondances sémantiques qui pourront éventuellement être utilisées dans l'approche intégrée, décrite au chapitre suivant, et dont l'objectif final est de répondre à des requêtes temporelles.

Dans la première partie de ce chapitre, nous présentons le cadre théorique du modèle Matching Distance, qui sera ensuite examiné sous l'angle des exigences définies par notre contexte, afin de justifier l'extension que nous nous proposons de lui ajouter. Nous définissons ensuite l'extension du modèle, où est décrite la méthode de segmentation et d'indexation des textes, la mesure de similarité entre les intervalles et la mesure pour de similarité pour les niveaux agrégés d'une hiérarchie. Finalement, nous présentons la fonction de rétablissement de liens sémantiques et les matrices de correspondances sémantiques qui seront intégrée dans l'approche géosémantique.

4. 2 Redéfinition du modèle de similarité Matching Distance

4.2.1 Modèle Matching Distance

Le modèle Matching Distance, conçu pour fournir aux systèmes d'information géographiques un outil pour l'intégration et la recherche d'informations, propose un cadre théorique pour évaluer la similarité sémantique entre des classes d'entités géospatiales. Alors que les modèles précédents se préoccupaient plutôt des propriétés géométriques de l'information géographique, le modèle Matching Distance en exploite les propriétés sémantiques et attribue une importance particulière à l'aspect cognitif de la similarité sémantique entre les entités spatiales.

La mesure de similarité du modèle Matching Distance associe des entités appartenant à la même ontologie (modèle Matching Distance) ou à des ontologies différentes (modèle Triple Matching Distance). Elle prend en considération les propriétés cognitives de la similarité et se base sur le modèle ratio de Tversky (Tversky, 1977) qui suggère que la similarité est une fonction des ensembles d'intersection et des différences de la forme suivante :

$$S(c_1, c_2) = \frac{f(C_1 \cap C_2)}{f(C_1 \cap C_2) + \alpha f(C_1 - C_2) + \beta f(C_2 - C_1)} \tag{4.1}$$

où c1 et c2 sont des concepts, C1 est l'ensemble des propriétés de c1 et C2 est l'ensemble des propriétés de c2, f est une fonction croissante monotone, autrement dit, pour tout couple d'élément (x1, x2) tels que x1≤x2, f(x1) ≤ f(x2). $\alpha \geq 0, \beta \geq 0$ sont des paramètres qui pondèrent les différences. La différence C1 - C2 correspond aux propriétés présentes dans C1 mais absentes de C2 et C2 – C1 correspond aux propriétés présentes dans C2 mais absentes de C1.

Dans le modèle Matching Distance, la similarité globale entre deux classes d'entités est une somme pondérée des similarités selon chaque type de propriétés (parties, attributs ou fonctions) (voir chapitre précédent), lesquelles sont données par la forme suivante du modèle ratio:

$$S(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha |C_1 - C_2| + (1-\alpha)|C_2 - C_1|} \tag{4.2}$$

Donc, dans le modèle Matching Distance, la fonction f présente dans l'équation du modèle ratio (équation 4.1) correspond à la cardinalité des ensembles C_1 et C_2 et $\beta = 1-\alpha$. Dans ce contexte, la cardinalité d'un ensemble correspond au nombre de propriétés constituant cet ensemble, par exemple le nombre de propriétés communes à c_1 et c_2.

Le modèle Triple Matching Distance englobe donc plusieurs aspects (contexte, propriétés des concepts, distance dans les graphes, similarité des voisinages dans les graphes et similarité lexicale) pouvant définir la similarité. Ainsi, il est plus complet que d'autres modèles qui se limitent à utiliser un seul aspect de la représentation des concepts, par exemple, ceux qui définissent la similarité uniquement par rapport aux propriétés des concepts (Tversky, 1977) ou à la distance dans les graphes (Rada, 1989 ; Jiang et Conrath, 1997).Cependant, bien que le modèle Matching Distance soit très complet, il doit être adapté pour rendre compte des caractéristiques du contexte de notre recherche, comme il en sera traité dans la section suivante ; cette adaptation aura également pour résultat de le

rendre plus flexible, c'est-à-dire pouvant prendre en compte une plus grande variété de types de propriétés.

4.2.2 Contexte du domaine forestier

Afin de développer notre approche de manière cohérente, nous avons préalablement identifié les exigences définies par notre contexte d'application, c'est-à-dire la forêt de Montmorency. Nous possédons quatre bases de données géospatiales multidimensionnelles pour chacun des quatre inventaires forestiers, effectués sur le même territoire, à des intervalles d'environ dix ans (1973, 1984, 1992 et 2001). Chaque inventaire représente ce territoire forestier par un ensemble d'unités spatiales de base, appelées peuplements. Un peuplement est une zone de la forêt qui est caractérisée par des propriétés homogènes en terme d'âge, de hauteur, de densité, d'essences, etc. Les peuplements sont par conséquent des entités spatiales décrites par une géométrie qui en donne la forme et la position dans l'espace et ils représentent les membres du plus fin niveau de la dimension spatiale. Au fil des inventaires, le découpage forestier évolue en raison de l'évolution naturelle de la forêt, des interventions humaines ou de l'évolution des modes d'acquisition, ce qui crée une hétérogénéité géométrique entre les différentes bases de données. Les inventaires comportent également leurs propres spécifications qui décrivent la manière dont les peuplements doivent être classifiés, par exemple des spécifications pour les classifications d'âge (jeune, mature, suranné, etc.). Les classifications d'essences, qui sont décrites par des textes, ont également été modifiées d'une année à l'autre, ceci ayant pour conséquence que, par exemple, le groupement d'essence *épinette* n'a pas la même signification d'un inventaire à l'autre (tableau 4.1) Cela crée une hétérogénéité sémantique entre les bases de données.

Inventaire	Essence	Descriptif
1973	Épinette	Épinette occupe plus de 50% de la surface
1984	Épinette	Épinette occupe au moins 75% de la surface
1992	Épinette	Épinette occupe plus de 84% de la surface

Tableau 4.1 : Évolution de la définition de l'essence épinette

Notre approche emploie une mesure de similarité dans le but de rétablir les liens sémantiques entre les membres des schémas des instances des bases de données. Les membres, qui sont les différentes zones du découpage forestier, soit les peuplements ainsi que les zones des niveaux supérieurs (figure 3.2), possèdent donc des propriétés qui prennent la forme de textes (pour les essences), d'intervalles numériques (par exemple pour les âges, la densité) et de mots (pour les fonctions, telles que réserves écologiques, production, bloc expérimental, etc.). De plus, comme le montre la figure 4.1, les membres sont liés par des relations (inclusion) dans un graphe qui représente la hiérarchie des dimensions spatiales.

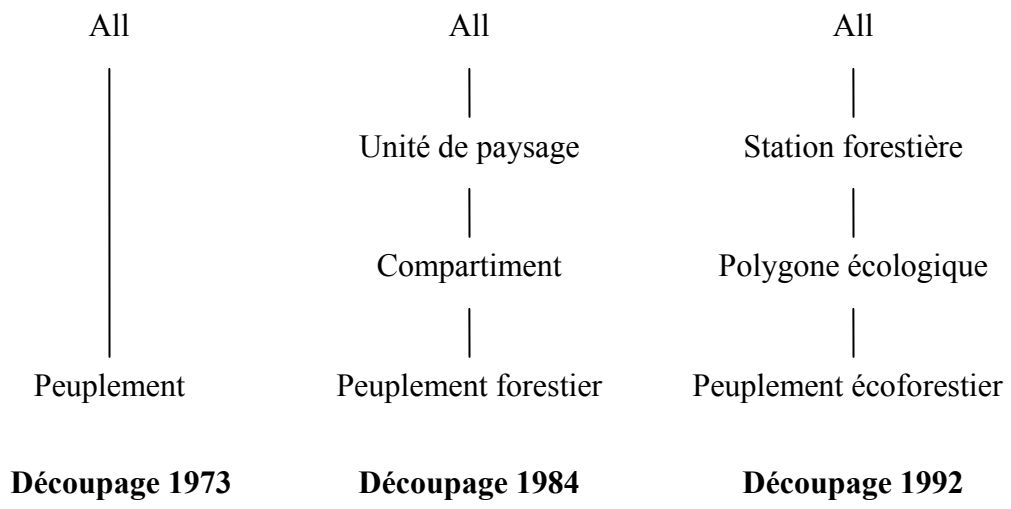

Figure 4.1: Hiérarchies des dimensions spatiales de la forêt de Montmorency

Les membres (instances) des schémas d'instances de notre application possèdent donc une représentation riche dont il est avantageux de tenir compte dans la mesure de la similarité afin d'obtenir les meilleurs résultats possibles. Nous considérons qu'une mesure de similarité qui ne tiendrait compte uniquement que d'un aspect de la représentation de ces instances (comme c'est le cas pour plusieurs mesures de similarité qui n'incorporent qu'un aspect de la représentation des concepts comparés, par exemple leur distance dans le graphe de l'ontologie) apparaît limitée dans l'optique d'une approche globale comme la nôtre où les instances à comparer sont définis sur plusieurs plans (propriétés, position dans la hiérarchie). Le modèle Triple Matching Distance, contrairement aux autres modèles décrits dans le chapitre précédent, offre cette possibilité de tenir compte de tous les aspects de la

représentation des concepts. Il offre aussi la possibilité de comparer les voisinages des concepts dans les graphes; ceci n'est pas pris en compte dans la plupart des modèles où la mesure de similarité est conçue pour quantifier la similarité sémantique entre des concepts qui appartiennent à la même ontologie et donc au même graphe, sauf bien sûr dans les modèles spécialement dédiés à la comparaison des graphes. La similarité entre les graphes et les parties de graphes est importante dans notre contexte puisque nous désirons évaluer la similarité entre des éléments qui appartiennent à des graphes différents, à savoir des schémas de dimension dont la structure est différente d'une époque à l'autre. Par contre, le niveau de définition du modèle Matching Distance est inapproprié dans le contexte où les concepts sont définis de manière plus complexe, comme c'est le cas dans le domaine forestier où les propriétés sont, par exemple, des textes ou des intervalles.

4.2.3 Limites du modèle Matching Distance

La mesure de similarité du modèle Matching Distance suggère que la fonction f du modèle ratio de Tversky corresponde à la cardinalité (| |) de l'ensemble d'intersection d'un type de propriété (équation 4.2). L'appartenance d'une propriété à l'ensemble d'intersection donne une réponse binaire, c'est-à-dire qu'une propriété fait partie ou ne fait pas partie de cet ensemble, mais elle ne peut pas y être incluse de manière partielle. Une propriété de la première entité fait partie de l'ensemble d'intersection de l'équation 4.2 si elle est identique ou si elle fait partie du même ensemble de synonymes qu'une propriété de la seconde entité; autrement elle en est exclue. Nous pouvons écrire ceci de la manière suivante :

Soit c_1 un concept et son ensemble de propriétés:

$$C_1 = \{\{C_{11}\},\{C_{12}\},\ldots\{C_{1j}\},\ldots\{C_{1m}\}\} \tag{4.3}$$

où $\{C_{1j}\}$ est un ensemble de synonymes de la propriété C_{1j} (incluant elle-même) et soit c_2 un second concept et son ensemble de propriétés:

$$C_2 = \{\{C_{21}\},\{C_{22}\},\ldots\{C_{2j}\},\ldots\{C_{2m}\}\} \tag{4.4}$$

où $\{C_{2j}\}$ est un ensemble de synonymes de la propriété C_{2j} (incluant elle-même). Alors pour une propriété C:

$$C \in C_1 \cap C_2 \text{ si } C \in C_1 \text{ et } C \in C_2 \tag{4.5}$$

Envisageons le cas simple où deux propriétés, telles que terrain de jeu et parc d'amusement sont des parties des entités stade et zone récréative. Supposons que ces propriétés ne sont pas considérées comme des synonymes dans l'ontologie. Ce couple de propriétés sera exclu de l'ensemble d'intersection et ne contribuera pas à augmenter la similarité entre les deux concepts même si elles ont des éléments communs sur le plan sémantique. Il est alors naturel de considérer que bien qu'elles ne soient pas des synonymes, ces propriétés doivent contribuer en partie à l'ensemble d'intersection. La mesure de similarité du modèle Matching Distance sous-estime donc la similarité de deux concepts, car elle ne considère pas le degré de ressemblance des propriétés lors de l'évaluation de l'intersection. Ce manque de précision peut devenir problématique dans le contexte où les entités sont très semblables entre elles, mais que le degré de ressemblance entre les propriétés est variable; le modèle Matching Distance sera alors impuissant à les distinguer. Pour une requête donnée, le modèle retourne alors un ensemble d'entités, qui bien que différentes, ne peuvent être classifiées selon leur degré de similarité avec l'entité de la requête. Un utilisateur ne possède pas suffisamment d'information pour prendre une décision.

D'autre part, les propriétés considérées dans le modèle Matching Distance sont représentées uniquement par des mots, ce qui exclut de pouvoir caractériser les entités par des formes plus complexes telles que des textes (ex. définition d'un concept par un texte) ou des intervalles numériques. Dans notre approche, ces limitations auraient beaucoup d'impact, car les éléments à comparer représentent des zones d'un découpage forestier décrites par des attributs très semblables ou qui diffèrent uniquement par les valeurs qu'ils prennent. Par exemple, dans l'inventaire forestier de 1984, les plus petites unités de découpage sont les peuplements forestiers qui sont décrits par les propriétés âge, hauteur, densité, perturbation, essence, fonction et dans l'inventaire forestier de 1992, les plus petites unités de découpage sont les peuplements écoforestiers qui diffèrent seulement par l'ajout de la propriété pente. Deux instances du concept peuplement forestier diffèrent par les valeurs des propriétés d'âge, de hauteur, de densité, etc. Notre contexte exige donc une mesure de

similarité qui soit la plus précise possible pour pouvoir les distinguer. Le fait que, dans notre contexte, les instances de concepts soient décrits par des propriétés complexes (c'est-à-dire composées de plusieurs éléments, par rapport à une propriété singulière comme un mot ou une valeur numérique unique) exige la redéfinition de la mesure de la similarité afin de pouvoir l'adapter à un contexte où ces propriétés ne sont pas seulement de type lexical.

4.2.4 Redéfinition du modèle Matching Distance

Sur le plan théorique, la redéfinition du modèle Matching Distance s'effectue selon la perspective générale où la similarité se fait au niveau des *concepts* dans l'objectif de rendre l'approche la plus générique possible. Toutefois, dans notre approche de rétablissement de liens sémantiques, la mesure de similarité sémantique redéfinie est employée pour comparer des instances, c'est-à-dire des membres des schémas d'instances des dimensions de différents cubes. En effet, les liens sémantiques doivent être rétablis au niveau des instances (par exemple entre des instances de peuplements P1 et P2) et non au niveau des concepts (c'est-à-dire entre les définitions générales des peuplements à différentes époques) car les requêtes temporelles sont posées au niveau des instances.

La figure 4.2 montre le processus global par lequel le modèle redéfini évalue la similarité sémantique entre les concepts. Les poids sont déterminés par le contexte, tel que défini à la section 3.3.2.5, selon le principe de variabilité ou de ressemblance. La similarité est d'abord évaluée entre les propriétés des concepts, soit au plus fin niveau de définition afin d'obtenir une précision maximale. Les propriétés peuvent prendre des formes complexes, telles que des textes ou des intervalles pour lesquels nous définirons des méthodes adaptées. La similarité entre les propriétés est ensuite incorporée dans l'expression de la similarité entre les concepts du plus fin niveau de la hiérarchie (le niveau détaillé) pour lesquels sont également évalué la similarité de voisinage et la similarité lexicale. Finalement, la similarité évaluée au niveau détaillé est incorporée dans la mesure de similarité pour les concepts des niveaux supérieurs de la hiérarchie.

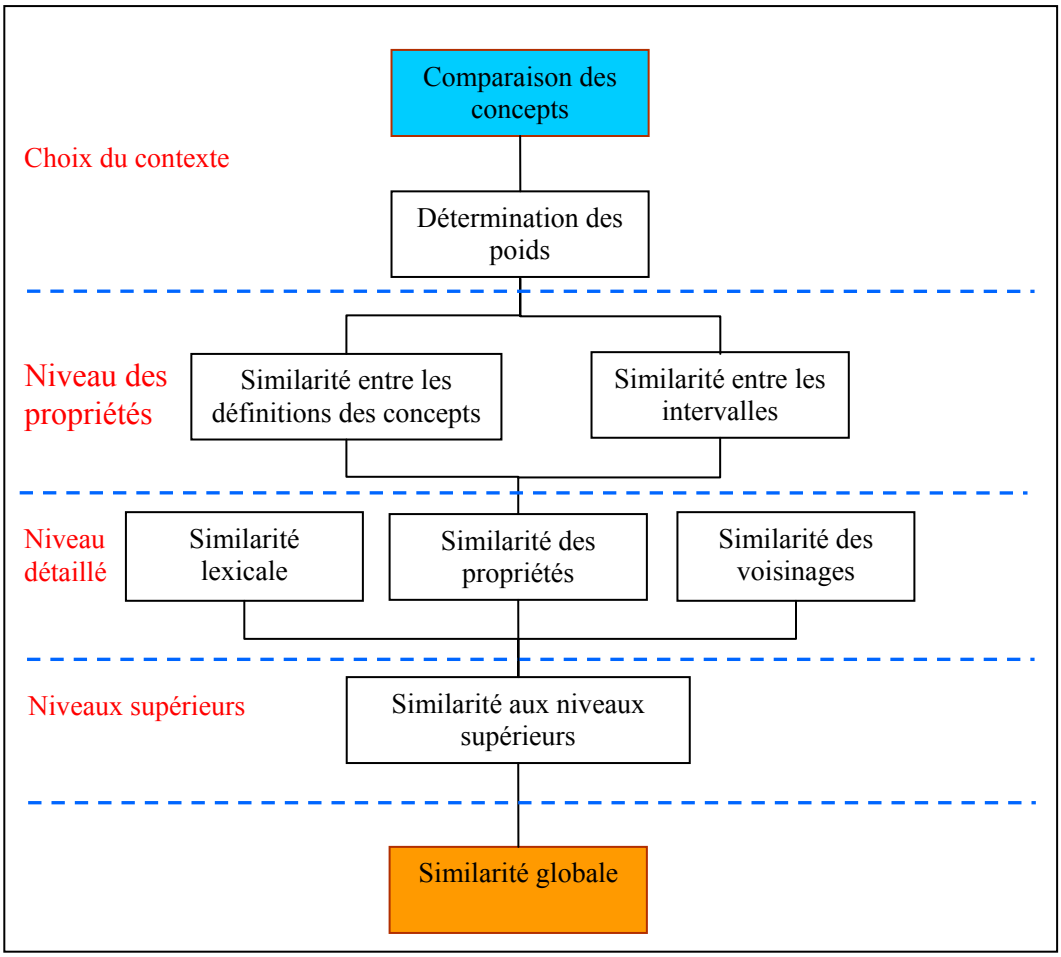

Figure 4.2: Processus d'évaluation de la similarité entre les concepts (modèle redéfini).

Afin de résoudre le problème de l'imprécision de la mesure de similarité du modèle Matching Distance, nous proposons de tenir compte de la similitude entre les propriétés lors de la détermination de la fonction f qui prend en argument l'ensemble des propriétés communes des concepts. La contribution d'un couple de propriétés à l'ensemble des propriétés communes doit être de l'ordre du pourcentage de similarité entre ces deux propriétés. Ainsi, si deux propriétés sont similaires à 45%, nous pouvons considérer que cela ajoute une valeur de 0.45 à l'ensemble d'intersection. Selon ce principe, la fonction de l'intersection peut être formalisée de la manière suivante:

Soit $C_{1t} = \{A_1, A_2, A_3, ...A_i, ...A_n\}_t$ l'ensemble des propriétés de type t de c_1 (t=attributs, parties, ou fonctions);

Soit $C_{2t} = \{B_1, B_2, B_3, ...B_i, ...B_m\}_t$ l'ensemble des propriétés de type t de c_2 (t=attributs, parties, ou fonctions);

$$f(C_{1t} \cap C_{2t}) = \sum_{i=1}^{n} \sum_{j=1}^{m} sim(A_i, B_j) \tag{4.6}$$

où sim(A_i, B_j) est une fonction de similarité qui dépend du type de données de A_i et B_j. Par exemple, si A_i et B_j sont des textes, le modèle doit être doté d'une mesure de similarité textuelle appropriée.

Puisque la différence entre deux ensembles E et F est définie par :

$E - F = E - (E \cap F)$ et que E ∩ F = F ∩ E

$$S(c_1, c_2) = \frac{f(C_1 \cap C_2)}{f(C_1 \cap C_2) + \alpha f(C_1 - C_2) + \beta f(C_2 - C_1)} \tag{4.7}$$

Nous aurons que:

$$f(C_1 - C_2) = |C_1| - \sum_{i=1}^{n} \sum_{j=1}^{m} sim(A_i, B_j)$$

$$f(C_2 - C_1) = |C_2| - \sum_{i=1}^{n} \sum_{j=1}^{m} sim(A_i, B_j) \tag{4.8}$$

Les différents types de propriétés envisagés dans notre approche sont des mots, des textes et des intervalles numériques. Toutefois, l'équation demeure générale et peut s'appliquer pour n'importe quel type de propriétés pour lequel il est possible de définir une mesure de similarité. Elle a pour avantage d'inclure la similarité entre les propriétés dans l'évaluation de la similarité entre les concepts, ce qui est fondamental dans des nombreuses situations où un plus fin niveau de définition des concepts doit être considéré.

Dans les sections suivantes de ce chapitre, nous définissons les différentes mesures de similarité adaptées aux différents types de propriétés qui peuvent être combinées au modèle redéfini.

4.2.4.1 Mesure de similarité entre les intervalles

Les concepts peuvent posséder des propriétés dont le domaine de valeur est déterminé par un intervalle. Ce peut être le cas de propriétés telles que l'âge, la hauteur, et la densité. Pour évaluer la similitude entre deux intervalles, nous devons employer une mesure de similarité qui prend en argument des valeurs continues plutôt que des valeurs discrètes, puisque dans ce cas un concept est déterminé par un ensemble infini de valeurs comprises dans un intervalle donné.

Pour choisir un modèle de mesure de similarité entre les intervalles, nous pouvons tenir compte du fait qu'un intervalle est un ensemble ordonné de points, E, tel que :

$$\exists t_1 \text{ tel que } \forall t \in \mathrm{E} \quad \mathrm{t}_1 \leq t$$

$$\exists t_2 \text{ tel que } \forall t \in \mathrm{E} \quad \mathrm{t}_2 \geq t \tag{4.9}$$

C'est-à-dire que t_1 et t_2 sont des bornes inférieures et supérieures.

Puisque les intervalles sont des ensembles, nous pouvons utiliser un modèle qui, a priori et dans sa forme de base, s'applique au cas où les concepts comparés sont décrits par des ensembles de propriétés ou de valeurs, ce qui est le cas des modèles vectoriels. Dans ce dernier modèle, l'ensemble des propriétés qui caractérisent les concepts forme un espace multidimensionnel où les axes sont les propriétés des concepts. Chaque concept est représenté par un vecteur v_c dans cet espace, où chaque composante v_i correspond à la fréquence d'une des propriétés dans le concept. Par exemple, si l'espace est formé des propriétés {texture, couleur, taille, forme, position, fonction} et qu'un concept est formé des propriétés {taille, forme, position}, alors le vecteur correspondant est (0, 0, 1, 1, 1, 0). L'intersection entre deux concepts, qui quantifie le commun entre les concepts, correspond au produit scalaire des vecteurs:

$$I(c_1, c_2) = \sum_{i=1}^{n} v_{1i} v_{2i} \tag{4.10}$$

Cette forme est utilisable si les propriétés des concepts sont discrètes, par exemple si une propriété est définie par un ensemble de valeurs discrètes, par exemple âge =

{0,1,2,….14,15}. Dans le cas de propriétés continues, par exemple si un concept est caractérisé par un intervalle numérique, par exemple âge= [0,15], l'intersection correspond à la limite de ce produit scalaire lorsque le nombre de propriétés tend vers l'infini et que l'écart entre les propriétés tend vers zéro. Si on a un ensemble X={v1, v2, v3,…,vn} et qu'on suppose que (vj-vi) → 0 et que n →∞ on obtient un intervalle $\Delta X = [v1, vn]$. Une fonction f qui associe à chaque point de l'intervalle la fréquence de cette valeur est une fonction de densité $\rho(r)$, qui donne la densité d'une propriété r :

$$I(c_1, c_2) = \int \rho_1(r)\rho_2(r)dr \tag{4.11}$$

La densité à l'intérieur d'un intervalle correspond à la distribution des données à l'intérieur des bornes de l'intervalle. Cette mesure est utilisée pour mesurer la similarité entre les molécules ; dans ce cas la densité correspond au nombre d'électrons entre un rayon r1 et un rayon r2. Dans le domaine forestier, la densité pour l'intervalle d'âge correspond, par exemple, à la densité d'arbres par hectare ayant entre 0 et 15 ans. Elle peut être une fonction continue ou discrète de l'âge. (Figure 4.3)

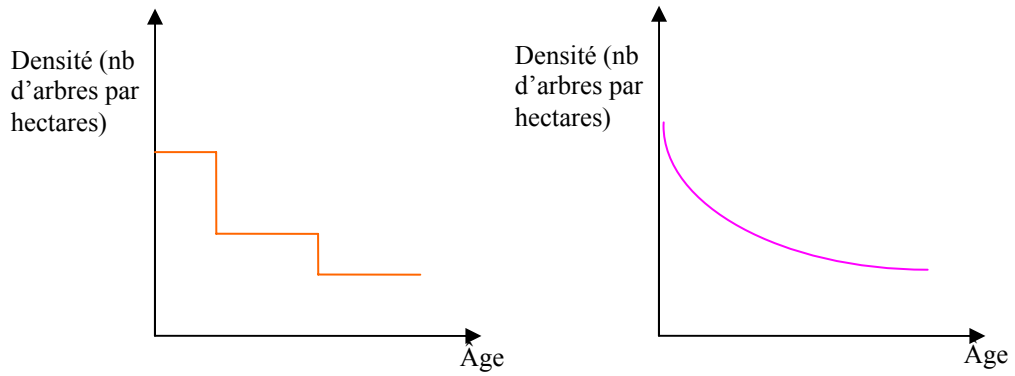

Figure 4.3 : Exemple de fonction de densité discrète (à gauche) ou continue (à droite).

La similarité est donnée par une fonction normalisée qui correspond à l'évaluation sur un domaine continu du cosinus de Salton (Salton, 1983):

$$sim(c_1, c_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \tag{4.12}$$

on obtient ainsi une mesure de similarité dont le domaine est compris entre 0 et 1 (Carbo-Dorca et Besalu, 1998):

$$sim(c_1, c_2) = \frac{I(c_1, c_2)}{\sqrt{I(c_1, c_1) I(c_2, c_2)}} \tag{4.13}$$

Nous employons cette mesure pour mesurer la similarité entre les intervalles, considérant que les vecteurs V_1 et V_2 sont les intervalles associés aux concepts c_1 et c_2. Soit A_i une propriété du concept c_1 telle que

$A_i = [a_o, a_1]$, avec $A_i \in D_k$ la dimension k et $\rho_a = \rho_a(r)$ est la densité à l'intérieur de A_i, et soit B_j une propriété de c_2 telle que

$B_j = [b_o, b_1]$, avec $B_j \in D_l$ la dimension l et $\rho_b = \rho_b(r)$ est la densité à l'intérieur de B_j

Si nom (D_k) ≠ nom (D_l), sim(A_i, B_j) = 0.

Si nom (D_k) = nom (D_l),

$$sim(A_i, B_j) = \frac{I(A_i, B_j)}{\sqrt{I(A_i, A_i) I(B_j, B_j)}} \tag{4.14}$$

Autrement dit, la similarité entre deux intervalles sera non nulle si ces intervalles donnent un domaine de valeur pour une même propriété. Dans le cas où une même propriété (correspondant à une même réalité) est désignée par des noms différents dans les deux concepts, il faut assumer que la correspondance entre les deux noms puisse être établie, par exemple, par un dictionnaire de synonymes. Dans le cas où la densité $\rho(r)$ est une fonction constante à l'intérieur d'un intervalle, c'est-à-dire que ρ pourra prendre la forme générale suivante:

$$\rho_a = \begin{cases} c_a & \text{si } a_0 \le r \le a_1 \\ 0 & \text{ailleurs} \end{cases} \quad ; \quad \rho_b = \begin{cases} c_b & \text{si } b_0 \le r \le b_1 \\ 0 & \text{ailleurs} \end{cases} \tag{4.15}$$

Il en découle directement que ces constantes peuvent être extraites de l'intégrale :

$$I(A_i, B_j) = \int \rho_a(r)\rho_b(r)dr = c_a c_b \int_{i1}^{i2} dr \tag{4.16}$$

où [i1, i2] est l'intersection entre les intervalles [a_0, a_1] et [b_0, b_1].

$$\text{Et } sim(A_i, B_j) = \frac{I(A_i, B_j)}{\sqrt{I(A_i, A_i)I(B_j, B_j)}} = \frac{c_a c_b \int_{i1}^{i2} dr}{\sqrt{c_a^2 \int_{a0}^{a1} dr * c_b^2 \int_{b0}^{b1} dr}} = \frac{\int_{i1}^{i2} dr}{\sqrt{\int_{a0}^{a1} dr * \int_{bo}^{b1} dr}} \tag{4.17}$$

Cela implique que lorsque la densité dans les intervalles est uniforme, la fonction de densité n'influence pas le calcul de la similarité. L'évaluation de la similarité entre les intervalles nécessite d'évaluer leur intersection [i1, i2], laquelle est déterminée par la relation entre les intervalles. Cette relation peut prendre plusieurs formes, un couple d'intervalles pouvant être caractérisé par sept relations de base, les relations d'Allen (Allen, 1983), résumées dans le tableau 4.2, où l'on compare l'intervalle A= [a_0, a_1] à l'intervalle B= [b_0, b_1]. Chaque relation, à l'exception de la relation « *égal* », est associée à une relation inverse (par exemple, l'inverse *de A inclus dans B* est *B inclus dans A*), par conséquent, il existe treize types de relations possibles entre deux intervalles.

Cette méthode pour l'évaluation de la similarité entre intervalles a pour avantage de permettre de mesurer la similarité entre les intervalles dans le cas où la distribution (la densité) est connue. Dans le cas où la fonction de densité est inconnue, la meilleure approximation consiste à supposer que la densité est uniforme à l'intérieur de l'intervalle, et on obtient une densité constante. Cette mesure de similarité entre les intervalles peut être incluse dans l'évaluation de la similarité entre les concepts.

Types de relation	notation	définition	illustration
A avant B	A<B	$a_0 < a_1 < b_0 < b_1$	a_0 a_1 b_0 b_1
A rencontre B	A m B	$a_0 < a_1 = b_0 < b_1$	a_0 a_1 b_0 b_1
A recouvre partiellement B	A o B	$a_0 < b_0 < a_1 < b_1$	a_0 a_1 b_0 b_1
A commence B	A s B	$b_0 < a_0 < a_1 = b_1$	a_0 a_1 b_0 b_1
A inclus dans B	A d B	$b_0 < a_0 < a_1 < b_1$	a_0 a_1 b_0 b_1
A termine B	A f B	$b_0 = a_0 < a_1 < b_1$	a_0 a_1 b_0 b_1
A égal B	A=B	$b_0 = a_0 < a_1 = b_1$	a_0 a_1 b_0 b_1

Tableau 4.2: Relations d'Allen

La section suivante montre le cas plus complexe de calcul de similarité où les propriétés des concepts sont des textes.

4.2.4.2 Mesure de similarité entre les définitions des concepts

Dans les systèmes de gestions de base de données (SGBD), le volume des données structurées représente un faible pourcentage par rapport aux données stockées sous forme de texte (Lefebvre, 2000). Celles-ci sont de riches sources d'information qui sont cependant difficilement accessibles car elles sont peu structurées. Les recherches dans le domaine des bases de données textuelles sont par conséquent primordiales et ont donné lieu à l'élaboration de techniques qui permettent d'accéder à l'information contenue dans les textes. Ces techniques sont l'objet des systèmes de recherche d'information. L'objectif visé par les systèmes de recherche d'information est de rendre opérationnelles les requêtes textuelles, c'est-à-dire de pouvoir sélectionner, parmi une collection de documents, les documents dont le contenu coïncide le plus possible avec un énoncé choisi par l'utilisateur. En règle générale, les composantes du système de recherche d'information sont un processus d'indexation, un processus d'appariement, un outil de prise de décision et un outil de jugement permettant de reformuler une requête (Amini, 2001).

Un système d'information a pour fondement une collection de documents, appelée le corpus, et son existence est motivée par un besoin en information formulé par une requête textuelle. Un processus de représentation des documents et de la requête textuelle est également nécessaire pour pouvoir précéder à la comparaison entre les documents ou entre un document et une requête. La phase de représentation des documents exige de procéder préalablement à l'indexation documentaire laquelle permet d'identifier parmi les documents l'information nécessaire à leur repérage. Cette information peut également permettre de constituer les mots-clés des requêtes textuelles. L'indexation comprend trois phases : l'extraction des éléments d'information, la transposition des éléments extraits dans un langage formalisé, le langage documentaire, qui est formé d'un ensemble de segments normalisés ; finalement l'attribution d'un index à chaque terme. Les éléments d'information, également appelés *descripteurs*, peuvent prendre la forme de mots simple auxquels sont soustraits les mots vides, des lemmes ou des racines des mots extraits, de concepts pouvant représenter plusieurs mots, de N-grammes (représentation d'un texte par

une séquence de caractères), de groupe de mots ou de manière plus large de contextes (Baziz, 2005). L'indexation effectuée sur l'ensemble du corpus permet de former le lexique. Finalement, la comparaison entre documents et entre documents et requête se fait au moyen d'un processus d'appariement, qui emploie habituellement des modèles de similarité qui évaluent la concordance entre les documents de la collection et la requête ainsi qu'un processus de sélection (filtrage d'information) permettant de prendre une décision. Les modèles de similarité sont aussi utilisés pour comparer les documents entre eux. Un processus complet de recherche d'information implique aussi un mécanisme de reformulation de la requête qui aide à la redéfinition du besoin en information pour le rendre comparable aux éléments des bases de données textuelles.

Notre approche nécessite de pouvoir comparer les descriptions textuelles de deux bases de données afin de déterminer un degré de similarité quantitatif entre eux. À chaque base de données représentant une époque donnée est associé un corpus de textes décrivant les groupements d'essences forestières. Le besoin en information, habituellement défini par une requête dans les systèmes de recherche d'information, est alors exprimé par un premier document extrait d'un premier corpus. Les processus d'indexation, de formation du lexique et de représentation des textes sont effectués pour l'ensemble des corpus comme dans la recherche d'information traditionnelle. La comparaison entre les textes se limite à l'évaluation de la similarité au moyen d'un modèle adapté à la représentation des textes. Dans notre contexte, nous n'effectuons pas de processus d'appariement et de filtrage par seuil puisque notre objectif est de quantifier la similarité entre tous les textes, et non d'obtenir des paires de textes correspondants l'un à l'autre. Le processus complet qui caractérise notre approche d'évaluation de la similarité des textes est illustré par la figure 4.4 et est décrit dans ce qui suit.

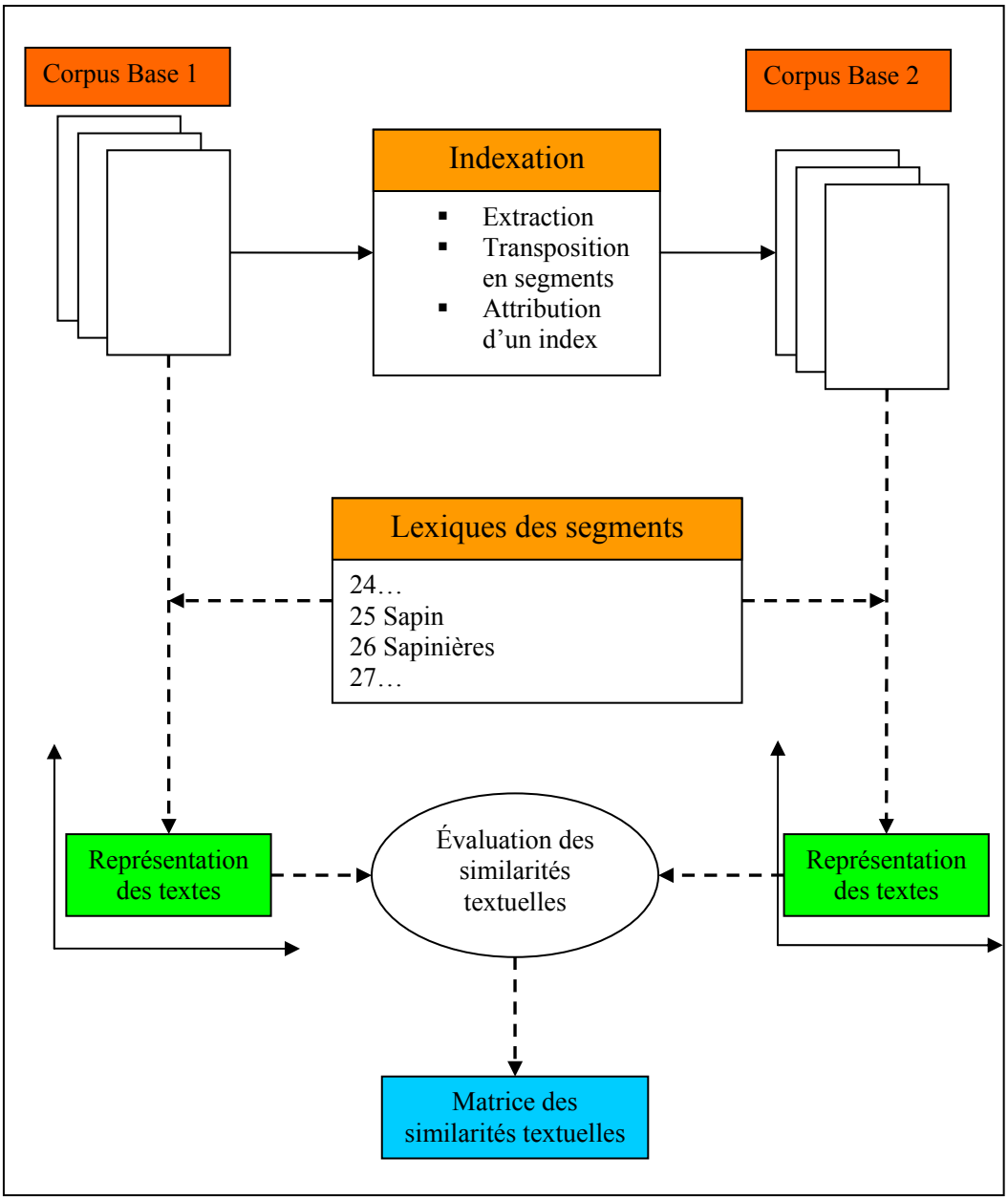

Figure 4.4: Processus d'évaluation des similarités textuelles

Dans notre contexte d'application forestière, les corpus sont constitués des spécifications sur les membres de la dimension *Essences*. Chaque groupement d'essences est décrit par un court texte qui décrit les conditions nécessaires pour qu'une zone de la forêt soit classifiée comme appartenant à ce groupe d'essence (tableau 4.3 et annexe 1A).

Code essence	Inven-taire	Descriptif	Type d'essence	Classe	Essence
BbS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau blanc occupe plus de 50% de la surface terrière de la partie feuillue. Dans ce peuplement, le sapin constitue plus de 50% de la surface terrière de la partie résineuse.	Mélangés	Bouleaux et résineux	Bouleaux blancs et sapins
BjS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau jaune occupe plus de 50% de la surface terrière de la partie feuillue. Dans ce peuplement, le sapin constitue plus de 50% de la surface terrière de la partie résineuse.	Mélangés	Bouleaux et résineux	Bouleaux jaunes et sapins
FiS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau blanc et les peupliers occupent, en proportion à peu près égales, plus de 50% de la surface terrière de la partie feuillue. Dans ce peuplement, le sapin constitue plus de 50% de la surface terrière de la partie résineuse.	Mélangés	Feuillus intolérants et résineux	Feuillus intolérants et sapins
EE	92	Peuplement où les résineux représentent 75% et plus de la surface terrière totale et où l'épinette noire et/ou rouge occupent 75% et plus de celle de la partie résineuse. On donne alors au peuplement le nom de cette essence.	Résineux	Un résineux dominant	Épinettes

Tableau 4.3: Extrait des descriptifs des essences de 1992

Indexation : L'indexation dans son ensemble, c'est-à-dire le repérage des segments, la transposition dans un langage formalisé et l'attribution d'un index aux segments formalisés, peut être réalisé de manière manuelle ou automatique. Le choix de l'une ou l'autre des techniques repose souvent du domaine modélisé par les documents ou du volume du corpus. Pour des textes spécialisés, l'indexation exige généralement d'être réalisée manuellement ou semi automatiquement en raison de la nécessité de recourir aux connaissances d'un expert du domaine. L'indexation manuelle est également plus précise et permet d'éviter la problématique de l'ambiguïté dans la signification des mots. Celle-ci se manifeste quand un mot possède plusieurs sens qui dépendent du contexte dans lequel il est employé (ce qu'on appelle également un mot polysémique) (Manning, et Schütze, 1999). Dans le cas d'indexation automatique, on doit avoir recours à des méthodes de désambiguïsation afin de déterminer le sens juste d'un mot étant donné le contexte. Bien que l'indexation manuelle soit préférable pour conserver le maximum de sens des textes, l'indexation manuelle est fastidieuse et s'avère impossible à réaliser pour un corpus

volumineux, dans quel cas l'indexation automatique est plus appropriée. L'indexation automatique peut être réalisée selon deux approches: l'approche statistique, qui se base sur la distribution des mots dans le corpus (Fagan, 1989), ou l'approche linguistique, qui se base par exemple sur le repérage de structures discursives (Chevallet, Haddad, 2001), ou encore des approches qui combinent ces deux dernières (Najib et al, 1996 ; Morin, 1999; Simoni, 2000) La première approche exige un corpus suffisamment important pour pouvoir identifier une distribution statistique.

Dans le cadre de notre approche, nous avons choisi une méthode d'indexation manuelle, d'une part car le corpus est peu volumineux et d'autre part parce que les textes sont de courtes descriptions constituées d'au plus quatre phrases. Les méthodes d'indexation automatique s'avèrent peu efficaces dans notre contexte ; nous devons exclure d'emblée les méthodes statistiques qui requièrent un corpus suffisamment volumineux ce qui n'est pas notre cas ; les approches linguistiques sont plus précises que les méthodes statistiques, c'est-à-dire qu'elles s'approchent plus du résultat désiré puisqu'elles tendent à reproduire le jugement humain. Il n'en demeure pas moins qu'elles sont susceptibles de produire des erreurs de sens, dont les répercussions peuvent être acceptables pour des documents de taille importante, mais qui ont des impacts importants sur des documents de petite taille, comme dans notre cas. En effet, si une méthode produit deux erreurs dans un court texte de quatre phrases, cela peut altérer considérablement le sens de ce texte.

Le processus d'indexation réalisé dans le cadre de notre approche comprend les trois phases suivantes:

Extraction des éléments informatifs: cette phase consiste à segmenter le texte en mots ou groupes de mots porteurs d'information et pertinents pour décrire ceux-ci. Les mots ou groupes de mots sont extraits du texte sans être transformés. Lors de l'extraction, nous avons observé un certain nombre de règles intuitives afin de préserver le maximum de sens:

- Regroupement des noms avec leurs adjectifs lorsque nous devons conserver le sens de cette association, par exemple pour le segment suivant : *Feuillus d'essences intolérantes.*

- Élimination des déterminants et des mots qui se répètent trop souvent (et, ou, dont...) entre les mots ou groupe de mots retenus.

- Regroupement des verbes avec leur sujet et leur complément lorsque nous devons conserver le sens de cette association.

Une fois les textes segmentés en éléments informatifs, la phase suivante consiste à réécrire ces éléments sous une forme générale qui permettra de procéder à leur comparaison.

Transposition des segments dans un langage normalisé: les éléments extraits dans la phase précédente étant dans leur forme brute, il est nécessaire de les traduire dans une forme normalisée. Ce processus, appelé lemmatisation, consiste à réécrire les mots sous leur forme canonique : les noms sont réécrit dans leur forme masculine, les verbes à l'infinitif et les adjectifs au singulier et au masculin afin de rendre les segments comparables entre eux. Lors de cette phase sont également éliminés les mots vides, c'est-à-dire ceux qui sont non significatifs car trop commun (et, ou, dont...), ce qui concerne habituellement les articles et les déterminants, sauf s'ils sont pertinents pour donner un sens. Dans ce cas, ils constituent un groupe de mots mais ne sont jamais laissés seuls. Par exemple, dans le fragment de texte suivant:

Fi73= *Feuillus d'essences intolérantes : jeune peuplement mélangé, dont le bouleau à papier et/ou les peupliers…*

Nous éliminons *dont*, mais pas *et/ou* parce que le dont n'aurait pas de sens en terme de similarité (dans tous les textes il est clair que plusieurs essences font partie du groupement d'essence, le *dont* n'impliquant aucune signification pour les essences sauf qu'elles font partie du peuplement, ce qui n'a pas besoin d'être mentionné). Par contre le *et/ou* est important en terme de sens, parce qu'il indique une possibilité de présence. Les mots ou groupes de mots résultant de la transposition forment les **segments informatifs** propres au corpus. Dans notre contexte, nous avons des textes courts d'une ou deux phrases qui comportent des structures semblables. Nous avons surtout retenu des groupements de mots pour préserver le sens des textes. En effet, la méthode de représentation choisie ne permet pas de tenir compte de l'ordre des mots dans le texte ; cependant on peut choisir des

groupes de mots qui ensemble portent le sens du texte. Il est essentiel de faire un pré-traitement qui permet de conserver le maximum de sens tout en ne choisissant pas de groupes de mots trop longs sans quoi ces groupes de mots seront peu susceptibles de se répéter dans l'ensemble des textes, ce qui diminuerait trop la valeur de similarité. Par exemple, si on représente la première partie du texte suivant :

Fi73 = *Feuillus d'essences intolérantes : jeune peuplement mélangé, dont le bouleau à papier et/ou les peupliers, seuls ou accompagnés d'une proportion variable d'érable rouge, occupent plus de 50% de la surface terrière de la partie feuillue ; est également classifié comme tel un peuplement mûr où la partie feuillue est composée des mêmes essences dans une proportion à peu près égale.*

Par les segments :

1-*Feuillus d'essences intolérantes : jeune peuplement mélangé ;*

2-*Bouleau à papier et/ou les peupliers, seuls ou accompagnés d'une proportion variable d'érable rouge ;*

La probabilité de trouver le premier segment dans un autre texte est assez faible, mais il n'est pas improbable que l'on trouve le segment *feuillu d'essences intolérantes.* Il faut donc fragmenter encore le premier segment pour s'assurer de détecter la similarité de deux textes qui comprendraient *feuillus d'essences intolérantes*. Par contre, on ne peut isoler le mot *intolérant* qui est clairement associé, dans ce contexte, aux feuillus. Il faut donc, d'une part, choisir un juste milieu dans la longueur des segments de manière à conserver le sens et permettre d'autre part, de détecter des similarités entre les textes. Nous avons également dû introduire des règles de segmentation propre à notre contexte. Dans plusieurs textes, un pourcentage décrit la couverture d'une zone par une essence particulière. Le segment constitué regroupe le pourcentage, l'essence et la zone affectée. Ce groupe de mots constituant le segment est placé dans un ordre constant pour chaque texte, à savoir la forme *qualificatif*, *pourcentage*, *sujet* (la zone), *verbe* (occuper dans la plupart des cas), *complément* (essence). On ajoute ensuite le mot de l'essence de façon isolée pour éviter que seul un pourcentage variable puisse réduire la similarité à zéro. Par exemple, les segments

informatifs 1 extraits des textes suivants possèdent la forme *qualificatif-pourcentage-sujet-verbe-complément* :

E73=*Peuplement où l'épinette noire et/ou rouge occupent au moins 50 % de la surface terrière de la partie résineuse du peuplement*

Segments informatifs :

1- Au moins 50% de surface terrière résineuse occupée par épinette noire et/ou rouge ;

2- Épinette noire et/ou rouge.

E(E) 84= *L'épinette noire et/ou rouge occupent au moins 75% de la surface terrière de la partie résineuse du peuplement.*

Segments informatifs :

1- Au moins 75% de surface terrière résineuse occupée par épinette noire et/ou rouge ;

2- Épinette noire et/ou rouge.

Dans ce cas, les textes ne peuvent être représentés uniquement par le premier segment car ils seraient jugés comme complètement différents ; ils doivent également être représentés par un second segment pour l'essence épinette noire et/ou rouge. La dissociation des essences et de leur pourcentage associé entraînerait une perte de sens dans la segmentation. La dernière étape à réaliser afin de former le lexique consiste à identifier les segments informatifs:

Attribution d'un index aux segments informatifs: à chaque segment informatif est associé un index qui permet de l'identifier et de le repérer. Cet index sera également utilisé dans l'étape suivante qui est la représentation des textes. Le résultat de cette phase est une table qui contient une liste ordonnée des segments ainsi que leur index respectif, qui constitue le lexique, et l'occurrence de ces segments dans chaque texte des corpus. Le processus d'indexation réalisé est représenté par la figure 4.5.

Libellé	Description
Bouleaux blancs et sapins	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terr
Bouleaux jaunes et sapins	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terr
Épinettes	Peuplement où les résineux représentent 75% et plus de la surface terrière totale
Épinettes et sapins	Peuplement où les résineux représentent 75% et plus de la surface terrière totale
Feuillus intolérants et sapins	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terr
Pas d'essence	Pas d'essence
Sapins et bouleaux blancs	Peuplement mélangé où les résineux représentent de 50% à 75% de la surface te
Sapins et bouleaux jaunes	Peuplement mélangé où les résineux représentent de 50% à 75% de la surface te
Sapins et épinettes	Peuplement où les résineux représentent 75% et plus de la surface terrière totale
Sapins et feuillus intolérants	Peuplement mélangé où les résineux représentent de 50% à 75% de la surface te
Sapins et peupliers	Peuplement mélangé où les résineux représentent de 50% à 75% de la surface te

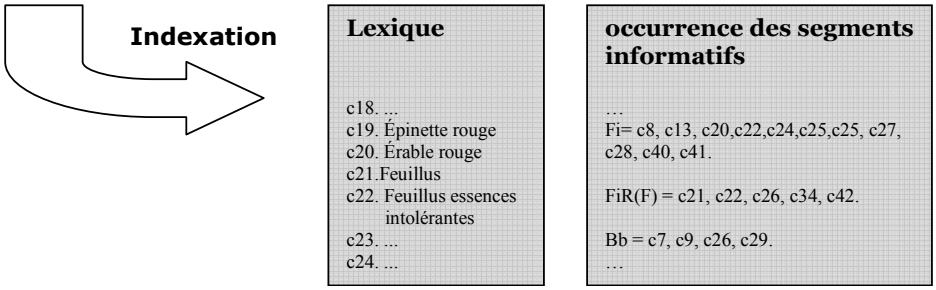

Figure 4.5: Processus d'indexation

Le lexique constitue la base du processus de représentation des documents qui est décrit ci-dessous.

Représentation des documents: La comparaison des documents s'effectue sur la base de leur représentation. Le modèle de représentation vectoriel, conceptuellement simple et intuitif en raison de l'efficacité de la métaphore spatiale, s'appuie sur un espace multidimensionnel afin de pouvoir y représenter les documents sous forme de vecteurs. Les dimensions de l'espace multidimensionnel correspondent aux termes présents dans le corpus, les termes étant les unités de base, des mots ou des groupes de mots, décrivant les textes (Salton, 1983). Dans notre contexte, les dimensions correspondent aux éléments du lexique et donc aux segments informatifs. Les composantes du vecteur représentant un document sont les fréquences de chaque segment informatif dans ce document. Dans notre approche, le modèle de représentation vectoriel est formalisé de la manière suivante:

Soit T_D le corpus global, qui regroupe l'ensemble des textes de tous les corpus associés à des bases particulières:

$$T_D = \{t_1, t_2, t_3, \ldots t_k\} \tag{4.19}$$

tel que $\forall t_i \in T_D, t_i \in D$, D une même dimension. Suite à la phase d'indexation, nous obtenons un ensemble C_D de segments informatifs indexés:

$$C_D = \{c_1, c_2, c_3, ...c_i, ...c_l\} \tag{4.20}$$

Les segments indexés et regroupés forment le lexique propre au corpus global et constituent le mode d'accès aux textes car ils définissent l'espace vectoriel dans lequel les textes sont représentés. Chaque texte t_i est décrit par un vecteur dans cet espace:

$$V(t_i) = \{v_{1i}, v_{2i}, v_{3i}, ...v_{ji}, ...v_{lj}\} \text{ avec } v_{ji} = f(c_j, t_i) \tag{4.21}$$

$f(c_j, t_i)$ représente la fréquence du segment c_j dans le texte t_i. Les vecteurs constitués lors de la phase de représentation permettent de procéder à la mesure de similarité.

Mesure de similarité entre les définitions de concepts (les textes)

Plusieurs mesures de similarité ont été conçues sur la base du modèle vectoriel, certaines d'entre elles étant booléennes, c'est-à-dire que les composantes du vecteur représentant un texte indiquent uniquement l'absence ou la présence (réponse binaire) d'un terme dans ce texte. Ces mesures sont basées sur la cardinalité de l'intersection entre les vecteurs X et Y: $|X \cap Y|$, qui donne le nombre de dimensions pour lesquelles les composantes des deux vecteurs sont non nulles. Par exemple le coefficient de Dice, qui normalise l'intersection en divisant par la somme des composantes non nulles pour les deux vecteurs, et qui donne une valeur comprise entre 0 et 1 (identique et dissimilaire):

$$sim(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{4.22}$$

Le coefficient de Jaccard normalise l'intersection en divisant par la cardinalité de l'union des vecteurs X et Y, c'est-à-dire par le nombre de composantes non nulles du vecteur X ou Y, ce qui donne une mesure de similarité plus faible que celle du coefficient de Dice:

$$sim(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{4.23}$$

Ces mesures, par contre, ne sont pas adaptées au cas où certains termes sont présents plus d'une fois, ce qui est le cas de certains textes décrivant les essences de la forêt de Montmorency. Par conséquent, l'usage d'une mesure de similarité se basant sur un modèle booléen semble peu approprié. Il existe également une mesure de similarité qui prend en compte la fréquence des termes dans un texte, la mesure du cosinus (Salton, 1983) qui est la seule mesure adaptée au modèle vectoriel n'opérant pas que sur des données binaires mais également sur des données quantitatives (Manning et Schütze, 1999) et qui plus utile dans notre contexte:

$$sim(X,Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}} \tag{4.24}$$

La mesure du cosinus vaut 1 lorsque les vecteurs sont alignés, ce qui correspond au cas où les textes sont considérés comme identiques, et 0 dans le cas où les vecteurs sont orthogonaux, ce qui correspond au cas où les textes ne partagent aucun segment informatif. En somme, plus les textes sont similaires, plus l'angle entre les vecteurs est faible et plus la similarité est élevée (s'approche de 1) (voir figure 4.6)

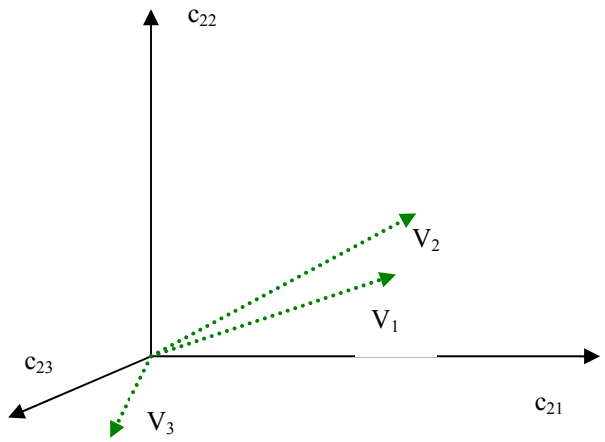

Figure 4.6: Similarité du cosinus. Les axes représentent les segments informatifs c_{21}, c_{22} et c_{23} du lexique. Le texte représenté par le vecteur V_2 est plus similaire au texte représenté par le vecteur V_2 qu'à celui représenté par le vecteur V_3.

Soit maintenant A_i une propriété de c_1 telle que

$$A_i = \{texte\}, \qquad \text{avec } A_i \in D_k \text{ la dimension } k$$

et B_j une propriété de c_2 telle que

$$B_i = \{texte\}, \qquad \text{avec } B_i \in D_l \text{ la dimension } l$$

Alors selon la représentation formalisée ci-dessus les textes sont représentés par des vecteurs dont les composantes sont les fréquences de chaque segment informatif du lexique dans le texte:

$$\begin{aligned} A_i &\rightarrow V(A_i) = \{v_{1i}, v_{2i}, v_{3i}, \ldots v_{ji}, \ldots v_{li}\} \\ B_j &\rightarrow V(B_j) = \{v_{1j}, v_{2j}, v_{3j}, \ldots v_{ij}, \ldots v_{lj}\} \end{aligned} \tag{4.25}$$

Nous avons ainsi constitué une table regroupant les vecteurs qui représentent chaque texte du corpus.

Pour le calcul de la similarité entre A_i et B_j, si les propriétés décrivent un membre de la même dimension, par exemple deux textes qui décrivent les groupements d'essences de la dimension *Essence*:

Si nom (D_k)= nom (D_l):

$$sim(A_i, B_j) = \cos(A_i, B_j) = \frac{\sum_{k=1}^{l} v_{ki} v_{kj}}{\sqrt{\sum_{k=1}^{l} v_{ki}^2 \sum_{k=1}^{l} v_{kj}^2}} \tag{4.26}$$

si nom(D_k)≠nom(D_l), $sim(A_i, B_j) = 0$ de telle sorte que dans le cas général, si les propriétés n'appartiennent pas à la même dimension, la similarité est automatiquement nulle puisque les textes ne décrivent pas des propriétés comparables.

Nous avons dans les deux sections précédentes, décrit des mesures de similarité qui s'appliquent dans le cas où les propriétés des concepts comparés sont caractérisées par des intervalles ou des textes. Ces mesures s'insèrent dans la similarité au niveau détaillé et permettent de calculer la similarité des concepts à un plus fin niveau de définition et pour des propriétés complexes. Un autre cas devant être défini est le cas des concepts des niveaux agrégés qui sont liés aux membres du niveau détaillé par des relations d'inclusion. La définition de la similarité des niveaux agrégés est explicitée dans la section suivante.

4.2.4.3 Mesure de similarité aux niveaux agrégés

Les schémas des dimensions spatiales des différents cubes sont structurés par une hiérarchie dont les niveaux concordent avec les différentes unités spatiales de gestion forestière. Le niveau le plus bas est le peuplement, les peuplements étant regroupés par compartiments (dans l'inventaire de 1984) ou par polygones écologiques (dans l'inventaire de 1992). Les compartiments forment les unités de paysage alors que les polygones écologiques sont rassemblés en stations forestières (figure 4.1). Le niveau le plus élevé (ALL) rassemble toute la surface de la forêt à toutes les époques où un inventaire fut effectué.

Les membres de chaque niveau de la hiérarchie sont liés par des relations d'inclusion, puisque, par exemple, les peuplements sont inclus dans les compartiments. Dans le domaine des ontologies, les relations d'inclusion sont dites relations *part-of*. Les parties des membres des niveaux supérieurs, tels que les polygones écologiques, sont donc les membres qui leur sont subordonnés par des relations *part-of* dans la hiérarchie (dans ce cas les peuplements écoforestiers).

Afin de prendre en compte cette situation, le modèle de similarité est redéfini par l'ajout d'une mesure de similarité pour des concepts se situant aux niveaux agrégés (supérieurs) d'une hiérarchie. Dans le modèle de similarité redéfini, comme dans le modèle Matching Distance, la similarité des parties aux niveaux supérieurs vérifie la ressemblance entre les constituants (parties) des concepts. Les parties des concepts étant elles-mêmes des concepts, le calcul de la similarité des parties aux niveaux agrégés de la hiérarchie est un processus récursif, la comparaison des ensembles de parties de chaque concept étant également une évaluation de la similarité. Soit $P(C_1) = \{p_{11}, p_{12}, p_{13}, ...p_{1i}, ...p_{1n}\}$ l'ensemble des parties qui composent le concept C_1, et soit $P(C_2) = \{p_{21}, p_{22}, p_{23}, ...p_{2i}, ...p_{2n}\}$ l'ensemble des parties qui composent le concept C_2.

La similarité des parties entre le concept C_1 et C_2 est donnée par:

$$S_p(C_1, C_2) = \frac{f(P(C_1), P(C_2))}{f(P(C_1), P(C_2)) + \alpha(C_1, C_2)(C_1 - C_2) + (1 - \alpha(C_1, C_2))(C_2 - C_1)} \tag{4.27}$$

Dans le modèle Matching Distance, la fonction f doit représenter l'ensemble des éléments communs (parties communes) entre C_1 et C_2. Puisque ces éléments sont eux-mêmes des concepts, le commun, ou l'intersection, entre les parties de C_1 et C_2 (la fonction f) doit être estimé par la ressemblance entre ces ensembles, soit par la somme des similarités entre les concepts les plus similaires. De plus, afin de respecter la règle selon laquelle l'intersection entre deux ensembles A et B est inférieure ou égale au plus petit des ensembles comparés :

$$\begin{aligned} &Si\ A \leq B,\ A \cap B \leq A \\ &Si\ B \leq A,\ A \cap B \leq B \end{aligned} \tag{4.28}$$

la fonction f doit avoir pour image l'ensemble des valeurs comprises entre zéro et la cardinalité du plus petit ensemble entre $P(C_1)$ et $P(C_2)$:

$$f:[0,1]\times[0,1]\rightarrow[0,\min\{|P(C_1)|,|P(C_2)|\}] \tag{4.29}$$

Finalement, la fonction f sera donnée par :

$$f(P(C_1)\cap P(C_2))=\begin{cases}\sum_{k=1}^{|P(C_1)|}\max\left[S_g(p_{1k},p_{2i})\right] & \text{si } |P(C_1)|\leq|P(C_2)| \quad ,i\in[1,|P(C_2)|]\\ \sum_{k=1}^{|P(C_2)|}\max\left[S_g(p_{1i},p_{2k})\right] & \text{si } |P(C_1)|>|P(C_2)| \quad ,i\in[1,|P(C_1)|]\end{cases} \tag{4.30}$$

Ce qui implique que si, par exemple, le concept C_1 possède moins de parties que le concept C_2, la fonction f sera donnée par la somme des similarités les plus élevées entre les parties de C_1 par rapport à celles de C_2 et réciproquement, si le concept C_2 possède moins de parties que le concept C_1, la fonction f sera donnée par la somme des similarités les plus élevées entre les parties de C_2 par rapport à celles de C_1.

Par définition, la différence entre deux concepts est donnée par le nombre de parties du premier concept auquel on soustrait le commun entre les parties des deux concepts, donné par la fonction $f(P(C_1),P(C_2))$:

$$\begin{aligned}(C_1-C_2)&=|P(C_1)|-f(P(C_1),P(C_2))\\(C_2-C_1)&=|P(C_2)|-f(P(C_1),P(C_2))\end{aligned} \tag{4.31}$$

Par exemple, considérons le graphe des ontologies de la figure 4.7 et l'évaluation de la similarité des parties entre le compartiment 7 et le polygone 4, où P(compartiment7)={p12,p13,p14,p15} et P(polygone4)={p21,p22,p23}. Considérons également que la similarité entre ces peuplements qui constituent le compartiment 7 et le polygone écologique 4 soit donnée par le tableau 4.4. Le calcul prend chacun des trois peuplements appartenant à l'ensemble des parties du polygone 4, puisqu'il est le plus petit des deux ensembles de parties, et identifie la similarité maximale entre un de ces peuplements et l'ensemble des parties du compartiment.

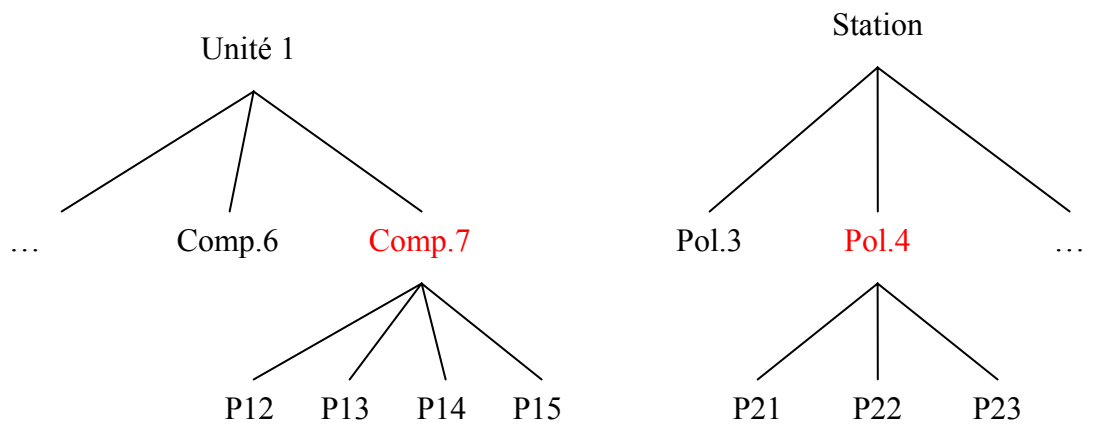

Figure 4.7: Similarité entre les membres des niveaux agrégés

P84/P92	21	22	23
12	0.5152	0.3252	0.0765
13	0.1234	0.4131	0.0965
14	0.3221	0.1623	0.3171
15	0.4453	0.3118	0.4418

Tableau 4.4 : **Exemple de valeur de similarité entre les peuplements de 1984 et 1992**

La fonction f équivaut à la somme de ces similarités maximales :

$$f(P(compartiment7), P(polygone4)) = 0.5152 + 0.4131 + 0.4418 = 1.3701$$

Par conséquent, puisque $\alpha(compartiment7, polygone4) = 1/2$, la similarité des parties entre le compartiment 7 et polygone 4 vaut :

$$S_p(C_1, C_2) = \frac{f(P(C_1), P(C_2))}{f(P(C_1), P(C_2)) + \alpha(C_1, C_2)(C_1 - C_2) + \alpha(C_1, C_2)(C_2 - C_1)}$$

$$S_p(C_1, C_2) = \frac{1.3701}{1.3701 + 1/2\,(4 - 1.3701) + (1 - 1/2)(3 - 1.3701)} = 0.3915$$

Cette mesure de similarité entre les parties de deux concepts peut être utilisée dans tous les cas où des relations d'inclusion lient des concepts. Si les concepts sont liés par d'autres types de relations, par exemple la relation de généralisation (relation *is-a* dans les ontologies), les enfants d'un concept dans le graphe de l'ontologie ne peuvent être considérées comme des parties de ce concept ; on devra considérer que pour les relations de généralisation, les concepts héritent des attributs des concepts du niveau supérieur et ce sera donc l'évaluation de la similarité des attributs qui sera un processus récursif.

Cette section complète la redéfinition du modèle de similarité sémantique qui constitue le cœur de notre approche pour rétablir les liens sémantiques entre deux versions de la base de données géospatiales.

4. 3 Rétablissement de liens sémantiques

Le modèle de similarité sémantique redéfini a pour fonction de peupler les matrices de relations sémantiques qui lient l'ensemble des membres du schéma d'instances d'une dimension d'une version de la base de données aux membres d'une seconde version. Plus précisément, une matrice de relation sémantique est calculée pour toutes les combinaisons de niveaux possibles (toujours entre deux bases différentes). Ceci permet non seulement d'identifier les relations entre les membres de même niveau mais également entre les membres de niveaux différents, afin de détecter des évolutions où des instances auraient été reclassifiées d'un niveau à l'autre. Dans cette section, nous définissons les éléments qui permettent de rétablir les liens sémantiques entre les schémas des instances des dimensions.

Ontologie évolutive. Les différentes versions de la structure multidimensionnelle sont représentées comme une ontologie évolutive, où les membres du schéma des instances d'une dimension sont des instances de concepts et les relations hiérarchiques entre les membres sont les relations entre les instances de concepts de l'ontologie. Une ontologie évolutive est constituée de plusieurs versions : O = {O1, O2, …Oi,... On}. Chaque version Oi est définie telle que : Oi = {Di, Ii, Ri, Ti}. Di représente l'ensemble des dimensions définies dans la structure multidimensionnelle. Ii= $\{c_1, c_2,\ldots, c_N\}$ représente l'ensemble des instances de concepts, lesquels correspondent aux membres des schémas des instances des dimensions. R_i est l'ensemble des relations entre les instances de concepts définissant la

structure de la hiérarchie des dimensions. T_i est un intervalle de temps valide pour la version i.

Définition des instances de concepts. Une instance de concept c d'une version O^i est définie de la manière suivante : c= {id_c, nom_c, P, D, L_D, O^i}, où id_c est l'identifiant de cette instance, nom_c est le nom de l'instance c, P est l'ensemble des propriétés (attributs, parties et fonctions) de c, D et L_D sont respectivement la dimension et le niveau hiérarchique auxquels appartient c. Les propriétés peuvent posséder des domaines de valeurs ou être définies par des textes.

Fonction de rétablissement de liens. Cette représentation de la structure sous forme d'une ontologie permet de définir un cadre théorique pour la recherche de relations sémantiques entre les instances des versions du cube par l'utilisation d'une fonction de rétablissement de liens sémantiques. Cette fonction quantifie, au moyen d'une mesure de similarité sémantique S_g, une relation de similarité entre deux instances. Celle-ci est une fonction composée qui se distingue selon le niveau de hiérarchie considéré, soit le niveau détaillé (plus fin niveau de granularité dans la hiérarchie) ou un des niveaux agrégés, la fonction des niveaux agrégés prenant en argument celle du niveau détaillé. La fonction du niveau détaillé met en relation les instances c_1^i et c_2^j des niveaux détaillés L_d^i et L_d^j de deux versions i et j de l'ontologie:

$$f_d : S_{gd}(c_1^i, c_2^j) \text{ avec } \mathrm{c}_1^{\mathrm{i}} \in L_d^i \text{ et } \mathrm{c}_2^{\mathrm{j}} \in L_d^j \tag{4.32}$$

La fonction du premier niveau agrégé met en relation une instance du niveau détaillé avec une instance d'un niveau agrégé ou deux instances appartenant aux niveaux agrégés:

$$f_{agg\_1} : S_{g\_agg}(S_{gd}(P_1^i, P_2^j)) \text{ avec } \mathrm{P}_1^{\mathrm{i}} = \begin{cases} \text{parties de } \mathrm{c}_1^{\mathrm{i}} & \text{si } \mathrm{c}_1^{\mathrm{i}} \in L_{agg}^i \\ c_1^i & \text{si } \mathrm{c}_1^{\mathrm{i}} \in L_d^i \end{cases} \tag{4.33}$$

Nous pouvons généraliser la fonction de rétablissement de liens du niveau agrégé à un niveau arbitraire n en employant le principe récursif :

$$f_{agg\_n} : S_{g\_agg\_n}(S_{g\_agg\_n-1}(S_{g\_agg\_n-2}(...(S_{gd} P_1^i, P_2^j)))). \tag{4.34}$$

Matrice de correspondances sémantiques. La fonction de rétablissement de liens sémantiques permet de définir les matrices de correspondances sémantiques qui mettent en relation les instances de deux niveaux appartenant à deux versions O^i et O^j de l'ontologie. La nécessité de rechercher une relation sémantique entre tous les niveaux possibles est justifiée par le besoin de pouvoir identifier le changement de niveau d'une instance lors de l'évolution. Soit $H(D,O^i) = \{c_1^i, c_2^i, \ldots, c_k^i, \ldots c_n^i\}$ l'ensemble des n instances formant le niveau L_1 de dimension D de la version O^i et soit $H(D,O^j) = \{c_1^j, c_2^j, \ldots, c_l^j, \ldots c_m^j\}$ l'ensemble des m instances formant le niveau L_2 de la dimension D appartenant à la version O^j. La matrice de correspondances sémantiques M est définie par :

$$M(D,O^i,O^j)_{kl} = f(c_k^i, c_l^j) \tag{4.35}$$

où f est la fonction du niveau détaillée f_d si L_1 et L_2 sont des niveaux détaillés et f_{agg} si L_1 ou L_2 est un niveau agrégé.

Ces matrices constituent le résultat final de l'approche sémantique et seront intégrées dans la méthode de transformation matricielle décrite au chapitre 6.

4. 4 Conclusion

Sur le plan sémantique, l'approche que nous avons définie dans ce chapitre permet, suite à l'évolution des membres de la structure, de rétablir les liens sémantiques nécessaires à l'analyse temporelle des données (Bakillah et al., 2006). Ces relations sémantiques sont représentées dans des matrices, lesquelles pourront être réutilisées dans le processus de réponse aux requêtes temporelles dans le cube de données géospatiales. Ce processus permettra d'ajuster la réponse aux requêtes temporelles puisqu'il intègre l'évolution sémantique des instances. Le modèle de similarité redéfini, bien qu'appliqué, dans le cadre de nos recherches, au domaine forestier, n'en demeure pas moins applicable dans tout domaine où des entités spatiales sont décrites par des propriétés de nature quelconque. En effet, cette extension du modèle s'avère intéressante puisque, loin d'être fermée à d'autres éventualités en ce qui concerne la nature des propriétés des entités comparées, elle a l'avantage de pouvoir éventuellement intégrer d'autre types de fonction de similarité entre des propriété de nature quelconque.

D’une manière similaire, la méthode proposée dans le chapitre suivant pour rétablir les liens géométriques permettra également de produire une matrice de relations géométriques entre les instances pour ajuster la réponse à la requête.

Chapitre 5 : Résolution du problème de l'évolution de la structure géométrique

5. 1 Introduction

Les différentes versions d'un cube de données géospatiales peuvent contenir des données cartographiques recueillies à différentes époques mais représentant le même territoire. L'évolution du découpage territorial fait en sorte que la comparaison entre les entités spatiales identifiées à chaque époque ne peut être réalisée directement. Cependant, afin de pouvoir traiter les requêtes spatio-temporelles portant sur ces entités, nous devons pouvoir les comparer pour identifier le pourcentage de chevauchement des polygones des différentes cartes. Pour accomplir cette tâche, il est nécessaire de ramener les entités spatiales des différentes époques à une représentation géométrique invariante dans le temps, ceci pouvant être réalisé au moyen de méthode d'indexation des données spatiales. Par la suite, il sera possible de constituer les matrices de correspondances géométriques d'une manière analogue à celle qui furent constituées pour le traitement de l'évolution sémantique. Dans la seconde partie de ce chapitre, c'est-à-dire suite à la présentation de l'approche géométrique, nous décrivons l'approche géosémantique qui permet d'intégrer les résultats du rétablissement des liens pour traiter les requêtes temporelles.

5. 2 Dimension spatiale et données géospatiales

Dans le contexte des bases de données multidimensionnelles, trois types de dimensions spatiales peuvent être répertoriés : la dimension spatiale non-géométrique, où les données qui sont dans la réalité possèdent une représentation spatiale mais ne possèdent pas de représentation géométriques dans la base de données (par exemple, un dimension formée par une hiérarchie de régions et de villes mais qui ne possède pas d'attribut géométrique dans la base de données), la dimension spatiale géométrique, où tous les niveaux ont une représentation géométrique, par exemple en étant associés à des polygones, et troisièmement la dimension mixte ou seulement certains niveaux sont représentés géométriquement. Les données spatiales peuvent prendre plusieurs formes. Dans les

systèmes d'information géographiques, l'information spatiale peut être conceptualisée par des champs, qui représentent des phénomènes continus dans un espace cartésien ou par des objets spatiaux, lesquels sont caractérisés par des attributs descriptifs et une géométrie (une forme et une position) (Zhang et Goodchild, 2002). Les objets spatiaux peuvent être de différents types, lesquels sont illustrés à la figure 5.1.

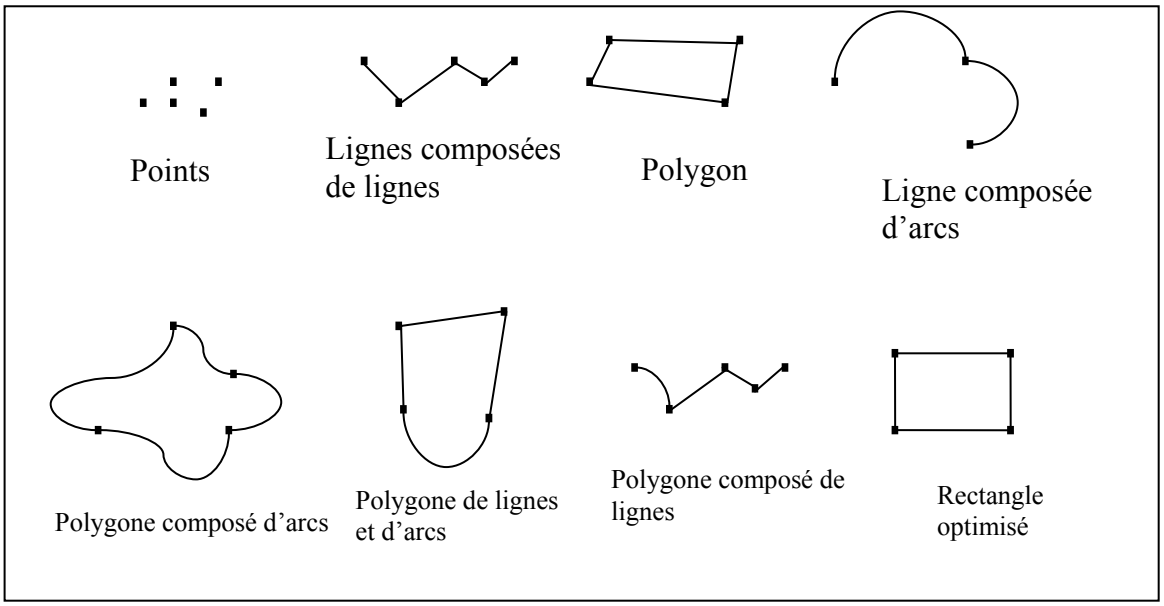

Figure 5.1: Types d'objets spatiaux (Rigaux et al., 2000)

Finalement, les données spatiales peuvent être représentées selon le mode matriciel (raster data), c'est-à-dire que les objets géométriques sont représentés dans une grille de cellules associées à des attributs ou des valeurs spectrales, ou le mode vectoriel, où les objets géométriques sont décrits par des coordonnées dans un système de référence donné. Le mode matriciel n'est pas particulièrement adapté pour représenter des objets individuels, ce qui convient plutôt au mode vectoriel. De plus, les données matricielles utilisent un important volume de stockage et présente une plus faible résolution que les données vectorielles. Par contre, le mode matriciel est fréquemment utilisé dans les outils d'analyse géospatiale car les opérations sur les données sont basées sur la segmentation du territoire en pixel et par agrégation de ces derniers on retrouve la surface initiale des polygones.

Dans les inventaires de la forêt Montmorency, les dimensions spatiales sont géométriques et chaque niveau est représenté par des objets spatiaux (polygones) selon le mode vectoriel. L'entité spatiale de base est le peuplement. Les peuplements des régions de la forêt où les propriétés descriptives, telles que l'âge, hauteur, densité, essences, etc. sont homogènes. Les peuplements qui présentent des propriétés semblables sont agrégés pour former les entités spatiales de niveaux agrégés, selon une hiérarchie qui a été modifiée à chaque inventaire. Les peuplements sont également associés à des mesures (leur surface, le volume ligneux, l'élévation, etc.).

Puisque le découpage du territoire forestier en peuplements est redéfini à chaque inventaire, le peuplement est un objet qui n'existe que pour une époque donnée et par conséquent dans une seule des versions du cube de données géospatiales représentant la forêt. Pour traiter les requêtes spatio-temporelles, il est nécessaire d'identifier leur pourcentage de chevauchement, ce qui sera rendu possible si les polygones des différentes époques peuvent être représentés par une même unité géométrique par une méthode de stockage et d'indexation de données spatiales appropriée. Avant de présenter notre approche, nous présentons un bref résumé des méthodes d'indexation.

5. 3 Méthodes d'indexation de données spatiales

L'augmentation du volume global de données géospatiales disponibles, par exemple avec la numérisation des cartes et le développement accru des techniques d'acquisition de données spatiales, fait en sorte que les besoins en terme de modélisation de données spatiales sont de plus en plus importants. De plus, le domaine des bases de données géospatiales est également investi par celui des bases de données temporelles, donnant naissance au domaine des bases de données spatio-temporelles nécessitant des méthodes d'indexation spatiales plus complexes, permettant, par exemple, d'adapter les structures de données spatiales existantes pour la représentation de phénomènes temporels (Tzouramanis et al, 2000). Les méthodes d'indexation des données spatiales ont pour rôle de supporter la représentation spatiale des objets et d'associer à ces représentations des opérations permettant de questionner la base de données. Ces méthodes se basent sur des structures de données spatiales qui doivent faciliter l'accès à ces données, tout en minimisant le temps de

réponse malgré le volume important des bases de données géospatiales. Il existe deux catégories de méthodes d'indexation de données spatiales (Rigaux et al, 2000): les structures basées sur l'espace, telles que les structures en grille et les arbres quaternaires (*quadtree*), où le partitionnement de l'espace est indépendant de la distribution des objets ; puis, à l'opposé, les structures basées sur la distribution (le regroupement) des objets spatiaux, dont l'exemple le plus connu est le R-Tree. Les structures en grille produisent un découpage de l'espace en cellules de taille uniforme (méthode *fixed grid*) ou en cellules de taille variable (méthode *grid file*) dans le cas où la distribution des objets est non uniforme ; elles ont pour inconvénient que le nombre de cellules formées par ce partitionnement peut croître très rapidement pour des grands volumes de données. Elles peuvent être appropriées si la distribution et la taille des objets spatiaux sont uniformes, par contre pour des objets de dimensions variables, la méthode des arbres quaternaires est plus appropriée puisque le découpage s'adapte à la densité de l'information spatiale. Ainsi, cela évite la présence d'un grand nombre de subdivisions vides, inutiles pour décrire la répartition des objets, ou la présence de cellules comportant au contraire trop d'information et qui devraient être divisées plus finement. Dans la méthode des arbres quaternaires, les zones créées lors du partitionnement sont représentées par les nœuds dans la structure hiérarchique de l'arbre quaternaire (un graphe entièrement connecté et acyclique, où chaque nœud possède quatre descendants). Il existe plusieurs variantes de l'arbre quaternaire qui varient selon le type de données spatiales (points, lignes, polygones, et autres objets de dimension agrégée, les règles régissant la division des nœuds et la résolution, qu'elle soit variable ou non (Hjaltason et Samet, 2002).

Les structures basées sur les objets, dont l'exemple typique est le R-tree, ont pour principe général de représenter des groupes d'objets dans l'espace. Les zones de l'espace englobant les groupes d'objets sont organisées hiérarchiquement selon les relations d'inclusions. Les arbres R ont été utilisés dans des méthodes d'indexation spatiotemporelles (Xu et al, 1990; Theodoridis et al, 1996 ; Nascimento et Silva, 1998; Nascimento et al, 1999). Ces méthodes sont plus efficaces que celles des arbres quaternaires pour effectuer certaines requêtes spatiales, par exemple les requêtes sur le voisinage d'un objet. Cependant, ils utilisent l'approximation du rectangle minimum englobant (RME) pour représenter les données spatiales, et donc l'approximation de la géométrie ne peut être raffinée, contrairement au

cas de la méthode de l'arbre quaternaire. Ils aussi pour inconvénient, par rapport à notre contexte, de permettre aux rectangles de se chevaucher, nuisant aux performances des requêtes, puisqu'il est possible d'emprunter plusieurs chemins dans l'arbre pour accéder aux objets, (car un objet peut être présent dans deux rectangles) ; mais surtout leur utilisation compliquerait le processus d'évaluation de la superposition. Par conséquent, elles s'avèrent inappropriées pour représenter et faire des opérations, en particulier d'intersection, sur des données surfaciques comme des polygones, à plus forte raison quand ceux-ci sont adjacents.

5. 4 Approche géométrique

L'objectif de l'approche géométrique est de quantifier la superposition des entités spatiales des différentes époques afin de constituer les matrices des correspondances géométriques, pour permettre de traiter les requêtes spatio-temporelles portant sur des entités spatiales évolutives. Afin de permettre la comparaison des entités spatiales, nous devons créer un mode de représentation uniforme pour toutes les époques. Pour cette tâche, la méthode d'indexation de l'arbre quaternaire (*quadtree*) s'avère, comme il en a été discuté dans la section précédente, la plus adaptée à l'objectif poursuivi ainsi qu'aux données spatiales choisies pour l'application de notre approche. L'application de la méthode des arbres quaternaires sur les données géométriques des différentes cartes permet de créer une représentation commune adéquate pour comparer les polygones. Bien que cette approche conduise à ne pas tenir compte de l'information sur la forme des polygones, ceux-ci sont reconstitués en terme des unités de base qui seront créées lors du découpage par la méthode quadtree, soit les cellules qui se situent au plus fin niveau de l'arbre quaternaire. En effet, il est important que les peuplements soient reconstitués puisque les entités géométriques appartenant à un découpage donné constituent la seule unité territoriale de base pertinente pour l'analyse spatio-temporelle dans le domaine forestier (Miquel, Bédard et al, 2001).

La figure 6.2 illustre l'approche géométrique adoptée. Les données géométriques (cartes) des différentes époques sont représentées à la même échelle et dans le même système de référence. Les données géométriques de chaque inventaire sont transformées en format ORACLE. Pour le partitionnement de l'espace avec la méthode de l'arbre quaternaire, nous

utilisons SPATIAL ORACLE, un module incluant des opérateurs et des algorithmes de stockage, d'accès et d'analyse des données géospatiales contenues dans une base de données ORACLE (Murray et al, 2003). Dans la phase suivante, l'identification des couples candidats permet de limiter le volume des opérations en sélectionnant les peuplements qui sont susceptibles de se chevaucher en tenant compte à la fois des relations géométriques et sémantiques. Le calcul des taux d'inclusion permet ensuite de peupler les matrices de correspondances géométriques qui seront utilisées pour traiter les requêtes spatio-temporelles.

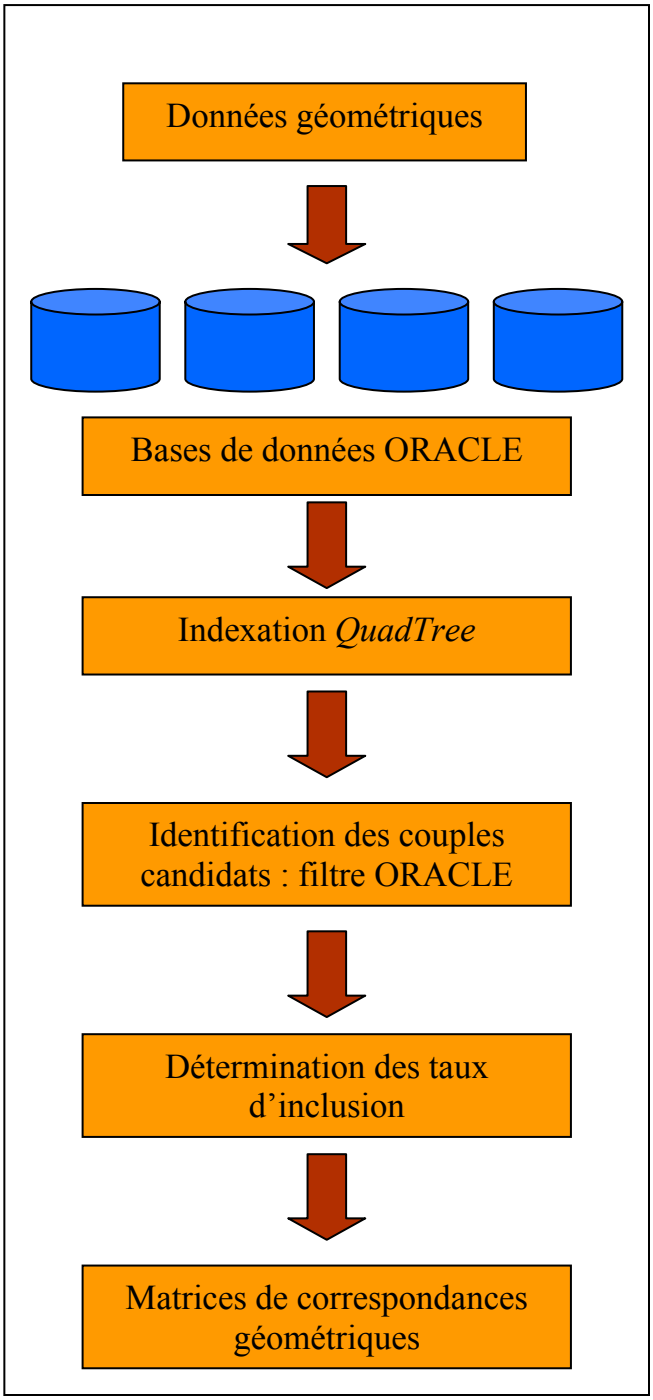

Figure 5.2: Approche de rétablissement de liens géométriques

Lors de la première phase, nous avons recréées sous format Oracle les tables de FAIT_ANNÉE des différents cubes de données géospatiales, avec les champs suivants:

```
CREATE TABLE FAIT_ANNÉE (
 ID_FAIT_ANNÉE                                NUMBER (10),
 ID_PEUPLEMENT_ANNÉE                          NUMBER (2),
 FK_DECOUPAGE_ANNÉE                           NUMBER (2),
 FK_AGE_DétAILLé_ANNÉE                        NUMBER (2),
 FK_ESSENCE_DétAILLée_ANNÉE                   NUMBER (4),
 FK_PERTURBATION_DétAILLée_ANNÉE              NUMBER (2),
 CODE_DENSITé                                 NUMBER(4),
 SUP_M                                        NUMBER (4),
 VOLUME                                       NUMBER (4),
 SUP_KM                                       NUMBER (4),
 GEOMETRIE                                    MDSYS.SDO_GEOMETRY);
```

Chaque table créée est associée à une table de métadonnées où nous avons inséré les valeurs pour le système de référence, les bornes inférieurs et supérieures pour les coordonnées du champ géométrie, et le taux d'incertitude (0.5m) à l'aide de la commande SQL, par exemple pour la table de FAIT_73:

```
insert into user_sdo_geom_metadata (table_name,column_name,srid,diminfo)
values('FAIT_ANNÉE','geometrie',32774,sdo_dim_array(sdo_dim_element('x',2
50000,260000,0.5),sdo_dim_element('y',520000,5300000,0.5)));
```

Chaque table de fait créée a été peuplée en insérant les coordonnées vectorielles, par exemple:

```
insert into Fait_92 (ID_PEUPLEMENT_92, GEOMETRIE) values (1,
MDsys.sdo_geometry(2003, Null, Null,mdsys.sdo_elem_info_array(1, 2003,1),
MDsys.sdo_ordinate_array(258706.796875, 5247636, 258706.4375 , 5247638.5,
258709.359375 ,5247651.5, 258711.09375 , 258706.796875));
```

La phase suivante consiste à employer la méthode *QuadTree* du module Oracle Spatial. Dans le processus d'indexation *QuadTree* (figure 5.3), l'espace des coordonnées subit une tessellation, c'est-à-dire que l'espace est divisé de manière hiérarchique et qu'à chaque objet spatial est assigné un ensemble de cellules issues de cette division récursive. Au départ, l'espace des coordonnées est considéré comme un rectangle et divisé en quadrants. Chaque quadrant engendré par la division est ensuite subdivisé à son tour en quadrant. Ce processus se poursuit tant que les cellules contiennent des éléments de la géométrie des objets spatiaux, ou jusqu'à l'atteinte d'un critère prédéterminé, par exemple la taille

minimale des quadrants. Les quadrants de taille minimale sont appelés les cellules. L'arbre quaternaire est un graphe entièrement connecté et acyclique permettent de représenter la structure des index spatiaux. Chaque nœud de l'arbre quaternaire représente un des quadrants, les nœuds du plus fin niveau représentant les cellules. Un noeud est associé à un code indiquant la position du quadrant auquel il est associé dans l'espace. Ce code est un nombre formé de *n* chiffres, *n* étant la profondeur locale de l'arbre. Par exemple, si le nœud se situe au niveau 3, il possède un code de trois chiffres. Les valeurs de ce code peuvent être 0, 1, 2, ou 3, chaque nombre correspondant respectivement aux quadrants nord-ouest, nord-est, sud-ouest et sud-est.

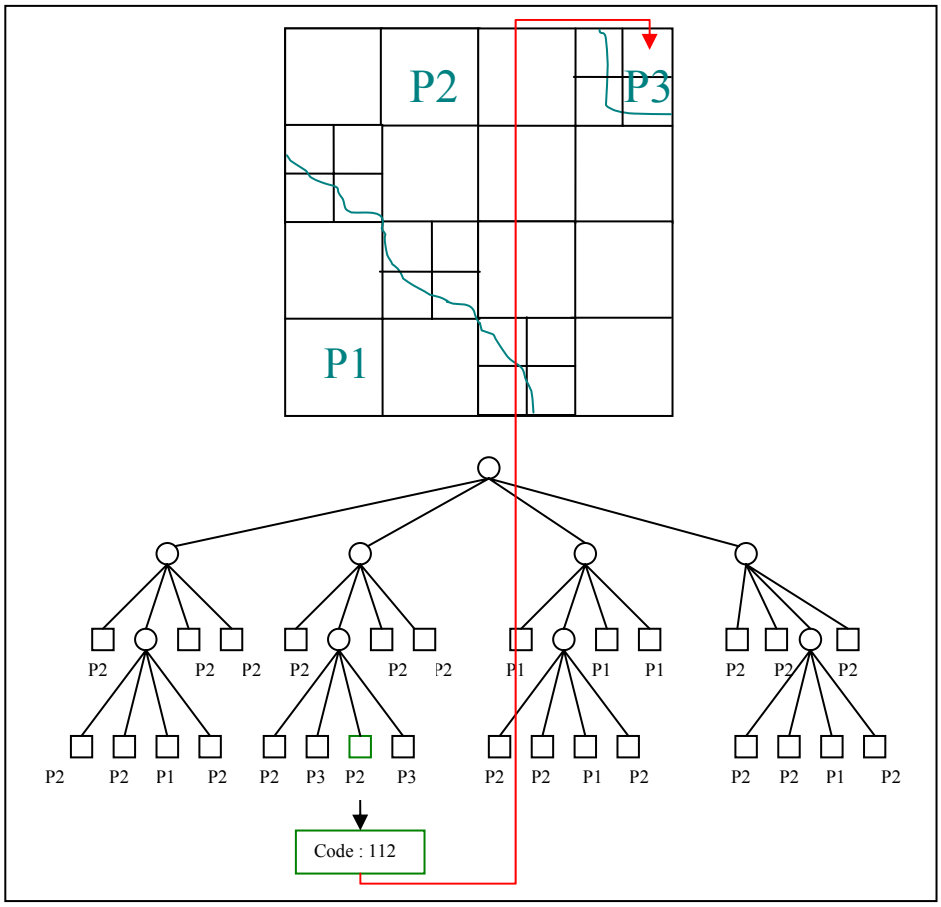

Figure 5.3: Partitionnement de l'espace selon la méthode de l'arbre quaternaire : la profondeur maximale a été fixée à n=3 et les cellules qui contiennent plusieurs polygones sont attribuées au polygone y occupant la plus grande surface.

Par exemple, le nœud représenté en vert est associé au code 112. Le processus de tessellation ORACLE est déterminé par deux paramètres : la résolution, SDO_LEVEL qui correspond aussi à la profondeur maximale de l'arbre, et le nombre maximal de cellules, SDO_NUMTILES. Le processus de tessellation peut être régi de deux manières : les cellules résultantes peuvent être de taille fixe ou variable. Dans le premier cas, seul le paramètre SDO_LEVEL est utilisé et la taille des cellules est donnée par la résolution, toutes les cellules ayant la même taille. Des cellules de taille variables sont obtenues lorsque seul le paramètre SDO_NUMTILES est défini. Lorsque les deux paramètres sont définis, l'indexation est dite hybride. L'indexation hybride est préférable lorsque la surface à indexer comprend beaucoup de polygones et que la taille des polygones est assez variable. En effet, les cartes de la forêt Montmorency sont formées d'un grand nombre de polygones qui sont de forme et étendues très disparates.

Les résultats de l'indexation sont stockés dans les tables SDOINDEX. Nous obtenons donc à la suite de cette procédure une table pour chaque inventaire indiquant les cellules qui forment chaque peuplement (figure 5.4), soit les tables FAIT_73_SDOINDEX, FAIT_84_SDOINDEX, et FAIT_92_SDOINDEX.

Id_peuplement_73	Index_cellules
5	0120311,0120312…
6	0120321,0120323…
7	0120320,0120322…
8	10001223,1000123…
9	100011,1000132…
10	1000133,10001312…
…	

Figure 5.4: Exemple de table d'indexation

Une fois les surfaces de tous les peuplements de chaque époque identifiés par leur ensemble de cellules respectives, chaque peuplement de chaque époque est mis en relation avec tous les autres peuplements d'une autre époque au moyen du taux d'inclusion, qui donne le pourcentage d'inclusion d'un peuplement P_j dans un peuplement P_i. Dans le cas où les bases de données sont constituées d'un grand nombre de peuplements, il est nécessaire d'optimiser l'algorithme en créant une fonction d'identification des couples de peuplements candidats, c'est-à-dire les couples de peuplements qui sont susceptibles de se chevaucher,

afin de ne pas calculer inutilement un taux d'inclusion qu'on sait être nul. La figure 5.5 montre la procédure de l'algorithme permettant de créer les matrices de correspondances géométriques à partir des tables d'indexation.

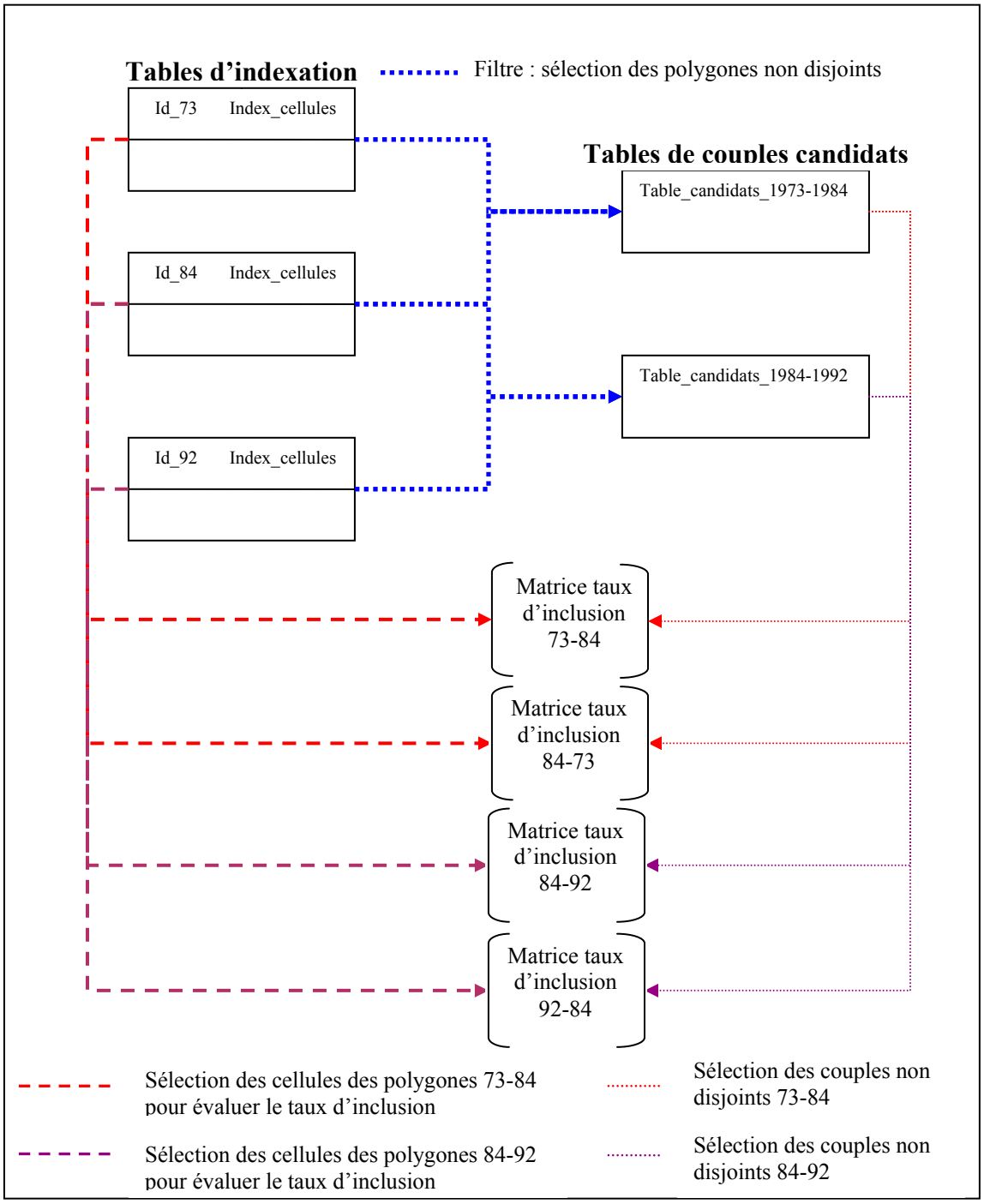

Figure 5.5: Algorithme de constitution des matrices de correspondances géométriques

Pour accomplir la sélection géométrique des couples de peuplements candidats et constituer les tables de couples de peuplements candidats, le module ORACLE offre déjà des fonctionnalités pouvant servir comme méthode de filtre spatial pouvant contribuer à la sélection des couples de peuplements candidats. Le filtre primaire SDO_FILTER permet de sélectionner parmi les polygones d'une première table, par exemple la table SDOINDEX_1992, ceux qui sont susceptibles d'être non disjoints d'un polygone requête P_i appartenant à une autre table, par exemple la table SDOINDEX_2002. Cette procédure est effectuée entre tous les polygones (peuplements) de toutes les tables d'indexation et permet de constituer plusieurs tables de stockage des couples de peuplements candidats : <*inventaire_cible, inventaire_source, peupl_id_source, id_peupl_voisins*> où chaque peuplement P_i de l'inventaire cible est associé à un ensemble de peuplements *peupl_voisins* $(P_i)= \{P_1, P_2, ..., P_j, ..., P_n\}$ d'un second inventaire source.

Le taux d'inclusion est évalué entre tous les peuplements présents dans les tables de couples candidats :

$$T(P_i, P_j) = \frac{S(P_i) \cap S(P_j)}{S(P_i)} \tag{5.1}$$

La représentation des peuplements sous forme de cellules permet de réécrire le taux d'inclusion de la manière suivante :

$$T(P_i, P_j) = \frac{\text{nombre de cellules commune à } \mathrm{P_i} \text{ et } \mathrm{P_j}}{\text{nombre de cellules formant } \mathrm{P_i}} \tag{5.2}$$

Les résultats du taux d'inclusion permettent de peupler la matrice de correspondance géométrique $M_g(G_1, G_2)$ qui met en relation les peuplements de l'inventaire cible G_1 avec ceux de l'inventaire géométrique source G_2. Les peuplements de l'ensemble cible sont identifiés aux lignes et ceux de l'ensemble source sont identifiés aux colonnes :

$$M_g(G_1,G_2) = \begin{array}{c} P_1 1 \\ P_1 2 \\ P_1 3 \\ \dots \\ P_1 n \end{array} \begin{pmatrix} T(P_1 1, P_2 1) & T(P_1 1, P_2 2) & \dots & \dots & T(P_1 1, P_2 m) \\ T(P_1 2, P_2 1) & & & & \\ \dots & & & & \\ \dots & & & & \\ T(P_1 n, P_2 1) & & & & T(P_1 n, P_2 m) \end{pmatrix} \quad (5.3)$$

Les matrices de correspondances géométriques sont donc constituées à partir des tables d'indexation par le champ commun qu'est l'identifiant des peuplements et qui permet de retracer les cellules qui constituent chaque peuplement dans les tables d'indexation.

La matrice de correspondance géométrique permet de transformer les mesures liées au peuplement d'un inventaire source (surface, volume ligneux…) dans la représentation cartographique de l'année cible par la méthode de transformation matricielle, qui sera décrite dans la section suivante. Ainsi, pour assurer la transformation des mesures d'une représentation vers l'autre, et vice-versa, deux matrices de correspondances géométriques sont nécessaire, les ensembles de peuplements de deux inventaires étant tour à tour ensemble cible ou ensemble source.

Dans la section suivante, nous décrivons comment les matrices de correspondances produites dans la phase de rétablissement de liens sémantiques (chapitre 4) et de liens géométriques sont intégrées dans l'approche géosémantique pour traiter les requêtes temporelles dans les cubes de données géospatiales.

5. 5 Approche géosémantique

L'approche géosémantique est conçue afin de permettre de traiter les requêtes temporelles dans des cubes de données géospatiales affectés par des évolutions sémantiques et géométriques en intégrant dans une méthode de transformation matricielle les résultats obtenus pour le rétablissement de liens sémantiques et géométriques. Plusieurs travaux ont traité l'évolution de la structure dans notre centre de recherche. Une des plus intéressantes approches a utilisé des méthodes d'agrégation pour résoudre les problèmes de l'évolution géométrique mais cette méthode ne considère pas le plus fin niveau de définition des peuplements sur le plan sémantique (Miguel, Bédard, Brisebois, 2002).

5.5.1 Méthode de transformation matricielle

La méthode de transformation matricielle se fonde sur la possibilité de considérer, sur le plan conceptuel, les cubes de données et les relations entre les membres du schéma des instances d'une dimension comme des matrices multidimensionnelles (Eder et Koncilia, 2001). Les matrices représentant les relations entre les membres de différentes versions permettent de lier les différentes versions d'un cube de données également représentées sous forme matricielle. La méthode permet de traiter tous les types d'évolution du schéma des instances, et, par conséquent, tous les types d'évolution du schéma de la dimension, puisque ces derniers peuvent toujours se traduire par des changements sur le schéma des instances (par exemple, la suppression d'un niveau hiérarchique se traduit, dans le schéma des instances, par la suppression de tous les membres du niveau à supprimer).

Définitions:

Une **version du cube** V_i est constituée d'un ensemble $D= \{d_1, d_2,..., d_n\}$ de n dimensions, chaque dimension d_i étant constituée d'un ensemble $M (d_i)= \{m_1, m_2....m_k\}$ de k membres (instances) et d'un ensemble de relations $R= \{r_1, r_2,..., r_i\}$ entre les membres, puis de faits, associant la valeur des mesures aux membres de ces dimensions. Une version du cube est également associée à intervalle de temps valide $[T_i, T_f]$ qui représente la durée pendant laquelle la totalité des membres et des relations composant la structure sont existants dans la réalité représentée.

La **matrice du cube** $C(V_i)$ est la conceptualisation d'une version du cube sous forme de matrice, possédant n dimensions qui correspondent aux n dimensions de la structure représentée. La taille de chaque dimension de la matrice étant donnée par $k(d_i)$, le nombre de membres formant la dimension correspondante d_i. La matrice du cube est peuplée par les cellules du cube, de telle sorte qu'à chaque combinaison de membres des différentes dimensions soit attribuée la valeur de la mesure correspondante dans le cube. Par exemple, une version du cube V_1 est constituée des deux dimensions suivantes : $D(V_1)= \{$*Découpage, Essence*$\}$ et possède les membres suivants pour le plus fin niveau : *M(Découpage)= {peuplement1, peuplement 2, peuplement 3}* et *M(Essence)= {peuplier, épinette}*. La matrice du cube C (V_1) a la forme suivante :

$$C(V_1) = \begin{array}{c} \\ p1 \\ p2 \\ p3 \end{array} \begin{array}{c} \begin{array}{cc} peuplier & épinette \end{array} \\ \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \end{pmatrix} \end{array}$$

Où f_{ij} représente le fait (valeur d'une mesure) associé à la combinaison des membres i et j.

La **matrice de transformation** *T (V_i, V_j)* est une matrice qui contient les relations entre les membres d'un niveau d'une dimension des versions du cube V_i et V_j et qui permet de transformer les faits issus d'une première version pour les représenter dans une seconde version. La matrice de transformation lie les membres de deux versions d'une même dimension, par exemple si la dimension *Découpage* existe dans deux versions mais possède une structure différente (des membres différents, des relations différentes…), l'élément ij de la matrice de transformation exprime la relation (sémantique ou géométrique) entre le membre i de la première version de la dimension (par exemple le polygone i) et le membre j de la seconde version de la dimension. La méthode de transformation matricielle permet de représenter les valeurs des mesures d'une version dans une seconde version par multiplication matricielle. La représentation *Rep (V_1, V_2)*, en une dimension, des faits de la version V_1 dans la version V_2 s'écrit de la manière suivante :

$$\mathrm{Rep}(V_1, V_2) = T(V_1, V_2, D_i) = C(V_1)T(V_1, V_2) \tag{5.4}$$

En généralisant ce principe pour représenter entièrement une version du cube dans une autre selon toutes les dimensions, nous obtenons:

$$\mathrm{Rep}(V_1, V_2) = T_{Dn}(T_{Dn-1}(...(T_{D2}(T_{D1}))))$$

5.5.2 Intégration de la méthode de transformation matricielle avec les approches sémantiques et géométriques

La méthode de transformation matricielle implique que les matrices de transformation sont peuplées de facteurs de poids qui constituent un lien quantitatif entre les membres de deux versions distinctes. Cependant, la nature de ces liens n'est pas mentionnée dans la mesure

où l'on conçoit qu'ils proviennent de la connaissance que l'on a, a priori des relations entre les membres, et du fait, également, que les évolutions de la structure sont crées par des opérateurs d'addition, de suppression, de fusion, etc. Or, l'étude des évolutions de la structure multidimensionnelle montre que l'évolution indirecte peut également affecter la structure et que dans ce cas, les approches de gestion de l'évolution proposées sont impuissantes puisque le rétablissement des liens, sur le plan conceptuel, est préalable à leur opérationnalisation via des fonctions de mapping. Le principe de l'approche géosémantique consiste à fusionner les méthodes de rétablissement des liens avec la méthode de transformation matricielle en considérant que les matrices de correspondances géométriques et sémantiques sont des matrices de transformation entre deux versions de la structure (figure 5.5)

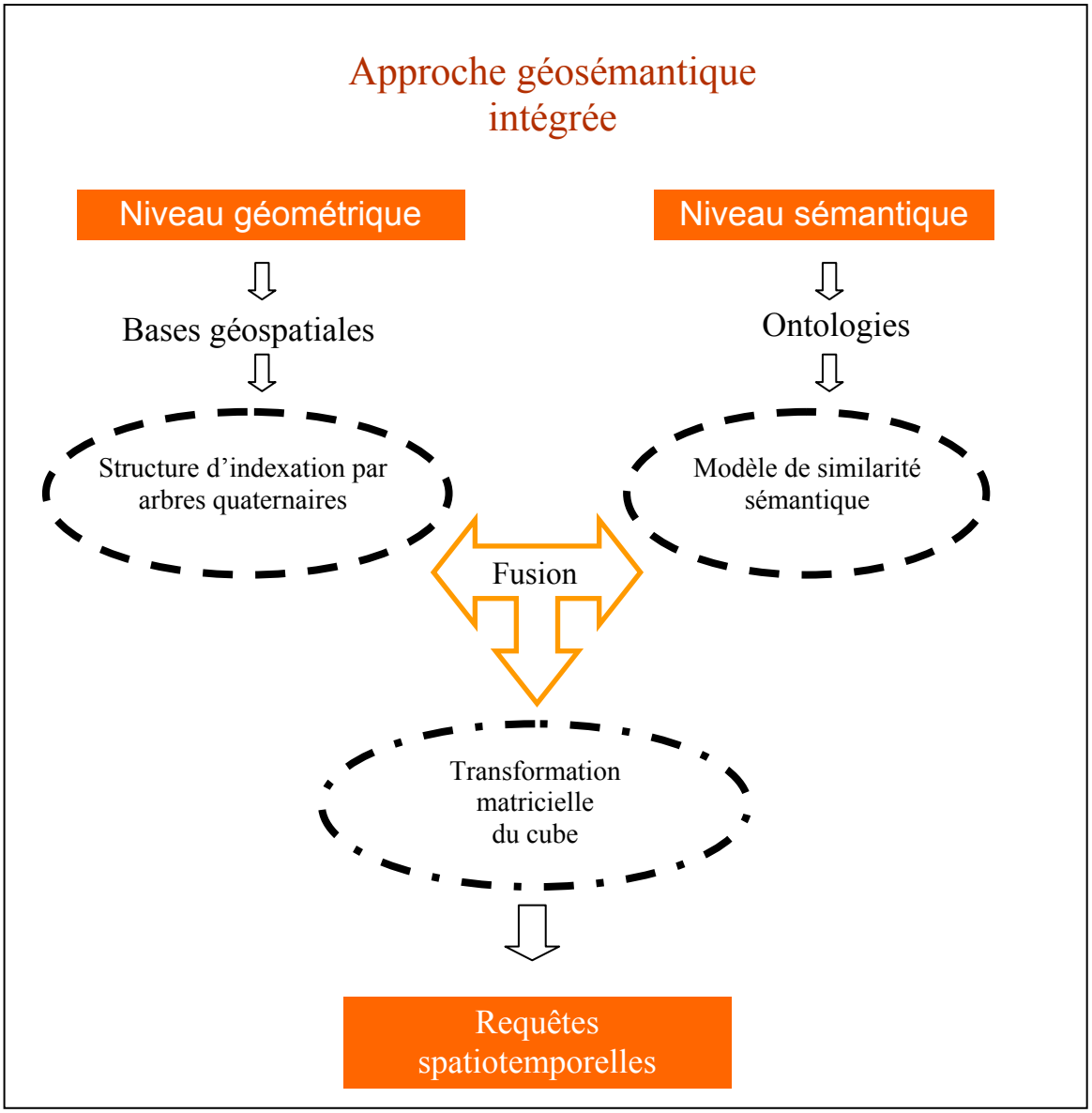

Figure 5.5 : Principe de l'approche géosémantique intégrée

La matrice des correspondances géométriques M_G est utilisée pour transformer les faits liés aux dimensions spatiales alors que la matrice de correspondances sémantiques M_S est utilisée pour transformer les faits liés aux dimensions dont les membres sont décrits par des attributs descriptifs de nature et de complexité variable, comme mentionné dans le chapitre 3. Les dimensions mixtes (spatiales et décrites par des attributs descriptifs) sont également traitées du fait que les matrices de correspondance géométriques peuvent être combinées pour former une matrice de correspondances globale, dite matrice de correspondances géosémantiques. La matrice de correspondances géosémantiques est donnée par:

$$M_{GS}(V_1,V_2)_{ij} = \left(M_G(V_1,V_2)\right)_{ij} * \left(M_S(V_1,V_2)\right)_{ij} \tag{5.5}$$

À l'aide des matrices de correspondance (géométrique, sémantique et géosémantique), la méthode de transformation matricielle permet de traiter des requêtes temporelles, c'est-à-dire des requêtes portant sur des membres ayant été modifiés au cours du temps, à condition de connaître les relations qui lient ces membres. Les requêtes sont traitées en fonction d'une version choisie par l'utilisateur, c'est-à-dire que si l'utilisateur choisi la structure V_1, toutes les données issues des versions concernées par la requêtes seront représentées par la méthode de transformation matricielle dans la version V_1, donnant ainsi à l'utilisateur une réponse cohérente avec les paramètres de la requête, là où auparavant il était même impossible de fournir une réponse à de telles requêtes. Une version du cube est concernée dans une requête si l'intervalle de temps valide de la version chevauche l'intervalle de temps mentionné dans la requête. Notre approche est suffisamment flexible puisqu'elle permet d'obtenir trois types de réponse, soit en utilisant comme matrice de transformation la matrice de correspondances sémantique, géométrique ou géosémantique, selon le type de dimension (spatiale, descriptive) considéré.

5. 6 Conclusion

Dans ce chapitre, nous avons présenté l'approche adoptée pour le rétablissement des liens géométriques entre les membres des schémas des instances des dimensions spatiales de différents cubes de données géospatiales représentant un même territoire à différentes époques. Parmi les différentes méthodes d'indexation, la méthode QuadTree s'avère appropriée privilégiée pour la comparaison de différentes cartes, étant déjà employé par exemple pour la comparaison d'images, car elle permet de créer une représentation commune pour différents ensemble d'objets géométriques, et ce d'une manière plus optimale, par exemple, que les structure en grilles qui créent une découpage régulier, c'est-à-dire non adapté à la distribution de l'information spatiale. Nous devons toutefois remarquer que l'approche de l'indexation QuadTree peut s'avérer moins précise et efficace dans le cas où les géométries des polygones ont peu évolué, comme c'est le cas pour quelques polygones entre 1984 et 1992. Dans ce cas, il est nécessaire de prévoir une méthode vectorielle qui serait plus précise (ex. appariement géométrique), le cadre de la

réalisation de ce mémoire n'offrant pas suffisamment de temps pour développer une approche alternative. La deuxième partie de ce chapitre, quant à elle, a été consacrée à la présentation de l'approche géosémantique qui fusionne les résultats des approches sémantiques et géométriques dans une méthode de transformation matricielle du cube. L'application de l'approche globale dans le contexte forestier est présentée dans le chapitre suivant.

Chapitre 6 : Application de l'approche géosémantique à un contexte forestier

6. 1 Introduction

L'implantation de notre approche a été réalisée afin d'en démontrer la validité. Dans une première partie, l'évaluation du modèle de similarité est présentée de manière notamment à montrer l'apport de la redéfinition du modèle par rapport au modèle original, le modèle Matching Distance. Nous discutons également les résultats obtenus concernant deux méthodes de détermination des poids attribués aux différentes similarités qui composent le modèle de similarité global. Dans une seconde partie, nous présentons l'application développée dans le domaine forestier, les différentes fonctionnalités implantées pour les requêtes spatio-temporelles ainsi que l'évaluation des résultats obtenus avec l'approche géosémantique.

6. 2 Évaluation du modèle de similarité

L'évaluation effectuée a pour objectif de montrer, d'une part, la justesse du comportement du modèle redéfini, et d'autre part, la performance accrue du modèle par rapport au modèle MD (Matching Distance).

Une simulation a été effectuée afin de valider le comportement du modèle de similarité sémantique redéfini. Le test consiste à évaluer la similarité entre une instance de référence choisie arbitrairement et avec un ensemble de 45 instances dont le pourcentage de propriétés communes avec l'instance de référence suit une fonction croissante linéaire, tel que la première instance de cet ensemble ne partage aucune propriétés avec l'instance de référence et que le dernier est identique à l'instance de référence. La figure 6.1 montre que la similarité suit une augmentation régulière par rapport à l'augmentation du pourcentage de propriétés communes et donc que le modèle redéfini se comporte selon la tendance attendue.

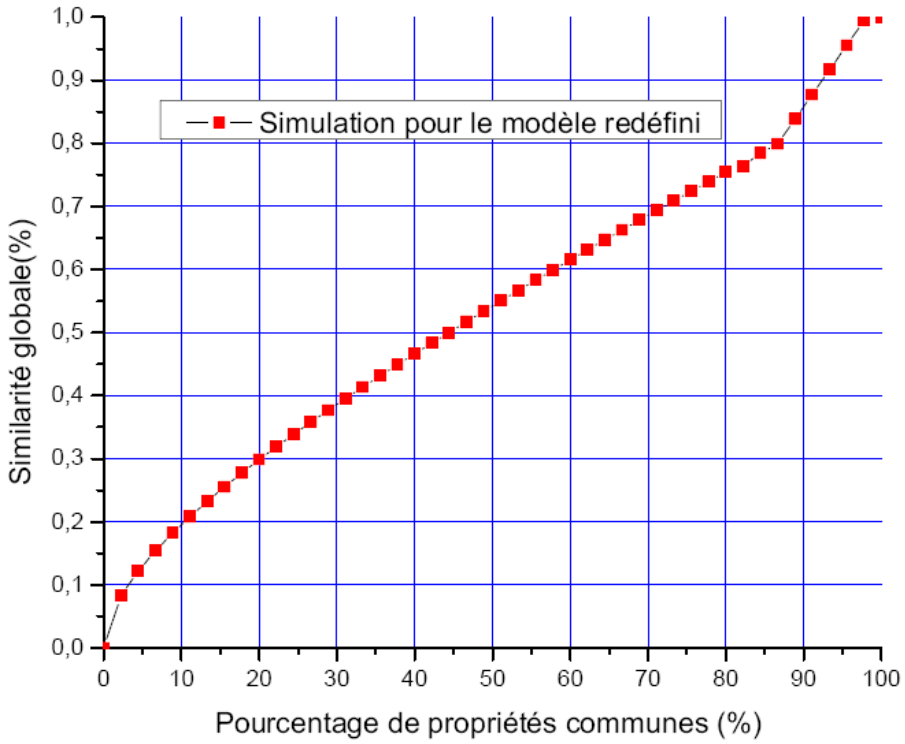

Fig. 6.1 : Test contrôle du modèle global de similarité redéfini

Afin d'évaluer les améliorations apportées au modèle de similarité MD, nous avons implanté ce dernier modèle afin de comparer les résultats de similarité obtenus avec le modèle MD et le modèle de similarité redéfini. La comparaison des résultats montre que pour toutes les classes de valeurs de similarité attendues, le modèle MD sous-estime la valeur de la similarité, car il rejette les propriétés qui sont partiellement similaires, alors que le modèle redéfini donne des valeurs de similarité plus rapprochées des valeurs attendues (tableau 6.1). Les catégories de valeurs de similarité attendues ont été identifiées manuellement à partir des spécifications des instances. Par exemple, alors qu'il fut jugé que les peuplements 23 (de l'inventaire de 1984) et 3 (de l'inventaire 1992) sont assez similaires, le modèle MD évalue leur similarité à 0.3712 alors que le modèle redéfini donne une valeur de 0.5874, soit une valeur plus près de ce qui pourrait être qualifié d'assez similaire.

Couples de peuplements (1984-1992)	Modèle MD	Modèle redéfini	Valeurs attendues (classes de similarité)
1818-3110	0.1031	0.4249	Assez similaires
1117-3110	0.0412	0.2528	Moyennement similaires
2395-3267	0.04125	0.1751	Moyennement similaires
1814-3075	0.0000	0.0438	Peu similaires

Tableau 6.1 : Comparaison de quelques valeurs obtenues pour le modèle redéfini et le modèle MD (annexe B, Tableau 1B et 2B)

L'efficacité du modèle MD et du modèle redéfini a été comparée par les mesures de précision et de rappel. Ces mesures sont des métriques qui sont couramment utilisées dans les systèmes de recherche d'information pour évaluer la performance de ces systèmes à identifier des couples de concepts ou de textes similaires. La précision évalue la proportion de liens correctement établis (I) par rapport à l'ensemble des liens établis (P) par le modèle, tandis que le rappel donne la proportion de liens correctement établis (I) par rapport aux liens attendus (R) (Rigaux et al., 2000):

$$\begin{aligned} \text{Précision} &= \frac{\text{nombre de liens correctement identifiés}}{\text{nombre de liens identifiés}} = \frac{\text{Card(I)}}{\text{Card(P)}} \\ \text{Rappel} &= \frac{\text{nombre de liens correctement identifiés}}{\text{nombre de liens attendus}} = \frac{\text{Card(I)}}{\text{Card(R)}} \end{aligned} \tag{6.1}$$

Les liens attendus, c'est-à-dire les couples d'instances qui sont considérés comme similaires dans la réalité, ont été déterminés manuellement afin de servir de référence lors de l'évaluation. Cette procédure a été effectuée en se basant sur les spécifications et des données cartographiques sur les entités spatiales comparées. L'évaluation a été effectuée à partir de 25 zones appartenant à l'inventaire de 1984 et de 77 zones appartenant à l'inventaire de 1992. Un nombre supérieur de zones candidates a été sélectionné en 1992 puisqu'à cette époque la même surface a été partitionnée en entités plus petites. Les liens attendus résultant de cette procédure sont présentés en annexe, soit au tableau 6B pour les instances utilisées pour le niveau détaillé et le tableau 7B pour les instances utilisées pour les niveaux agrégés. Les valeurs pour la cardinalité de l'ensemble de liens identifiés (P) sont obtenues en appliquant successivement des valeurs de seuil décroissantes aux valeurs de similarité, au moyen d'une fonction permettant de choisir manuellement la valeur du seuil à appliquer aux matrices de correspondances sémantiques. La fonction de seuil

génère, à partir d'une matrice de correspondances sémantiques, une matrice binaire en attribuant une valeur de 1 aux couples d'instances liés par une valeur de similarité supérieure ou égale au seuil et une valeur de 0 aux couples liés par une valeur de similarité inférieure au seuil. L'ensemble P regroupe alors tous les couples d'instances liés par une valeur de 1. L'ensemble I est un sous-ensemble de P et correspond aux couples qui se situent à la fois dans P et dans l'ensemble de référence R. Les figures 6.2 et 6.3 illustrent les résultats de la précision en fonction du rappel pour la fonction de similarité aux niveaux agrégés et au niveau détaillé

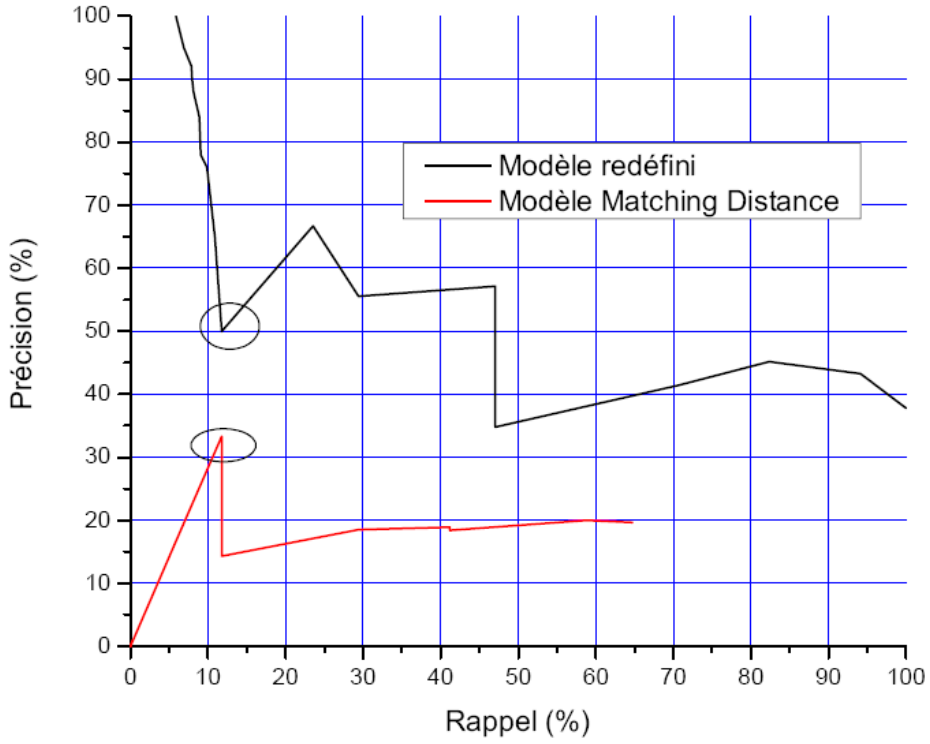

Figure 6.2: Comparaison des courbes de précision en fonction du rappel pour le modèle redéfini et le modèle MD pour les niveaux agrégés

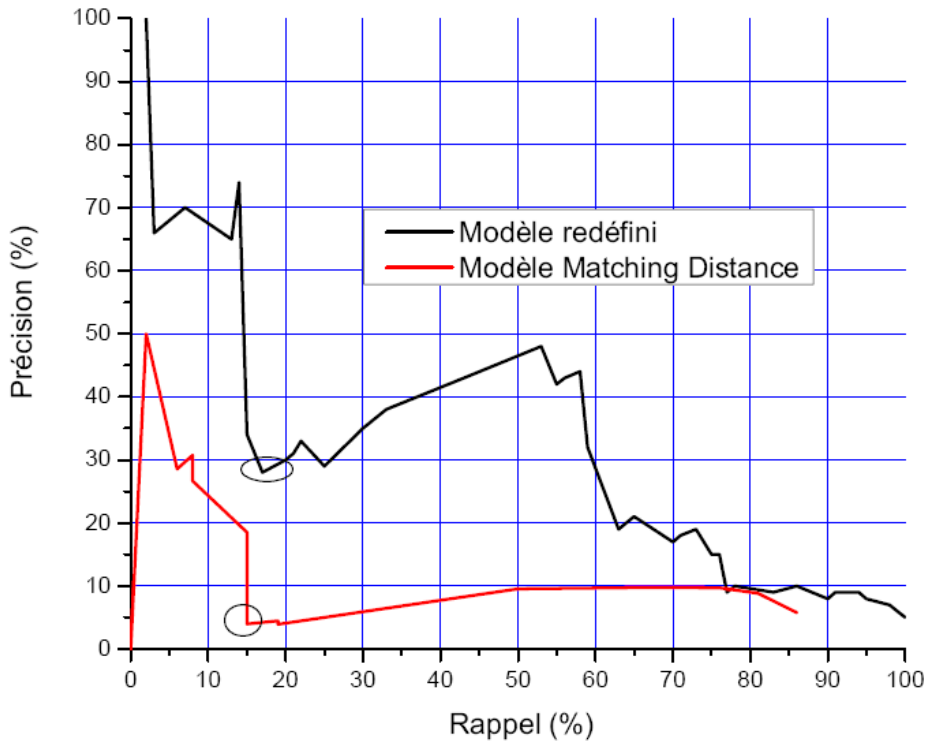

Figure 6.3 : Comparaison des courbes de précision en fonction du rappel pour le modèle redéfini et le modèle MD pour le niveau détaillé

Les résultats montrent que la performance du modèle redéfini est supérieure en tout point à celle du modèle MD, en particulier, l'écart entre les deux modèles étant plus important dans le cas de la similarité aux niveaux agrégés, ce qui montre la nécessité d'utiliser une mesure de similarité spécifiquement adaptée aux niveaux agrégés, particulièrement dans le cas ou les membres des niveaux agrégés ne possèdent pas de propriétés intrinsèques et sont uniquement définis par les concepts du niveau détaillé. En effet, dans ce cas, la similarité du modèle MD se réduit à la similarité entre les voisinages des concepts dans le graphe et à la similarité lexicale, alors que le modèle redéfini permet de considérer implicitement la similarité entre les propriétés dans la fonction récursive. Les courbes montrent des variations assez marquées (encerclées sur les graphiques), reflétant le regroupement des valeurs de similarité entre les instances. La distribution des valeurs de similarité n'est pas uniforme mais forme des regroupements, c'est-à-dire que la fréquence de certaines valeurs de similarité est beaucoup plus importante que pour d'autres. Ces regroupements sont causés par la variabilité limitée des propriétés (hauteur, densité, essences…) caractérisant

les peuplements. Ceci engendre une forte sensibilité des valeurs de rappel et de précision par rapport à de faibles variations du seuil pendant le test.

6. 3 Méthode de détermination des poids pour les différents types de similarités

Dans cette section, nous avons étudié l'impact des différentes méthodes de détermination des poids attribués aux différentes similarités (attributs, parties, fonctions) sur les valeurs de similarité obtenues. La méthode de détermination des poids est primordiale puisqu'elle est susceptible de modifier de manière significative les valeurs de similarité. Les poids des différentes similarités peuvent être déterminés manuellement, c'est-à-dire qu'un utilisateur familier au contexte peut choisir, par exemple, d'attribuer plus d'importance relative à la similarité des parties, par rapport à la similarité des attributs ou des fonctions. Cependant, un tel choix n'est pas toujours souhaitable ni possible, en particulier quand l'utilisateur ne possède aucun repère pour estimer des valeurs optimales pour ces poids. Une détermination arbitraire peut par conséquent s'avérer non représentative du contexte forestier et par conséquent affecter de manière imprévisible les valeurs de similarité. Les approches probabilistes de détermination des poids que sont la ressemblance et la variabilité permettent de calculer les poids en fonction du contexte, c'est-à-dire en fonction de la distribution statistique des propriétés des instances parmi un ensemble d'instances présentant un intérêt pour l'utilisateur. Nous avons testé les approches de la ressemblance et de la variabilité afin de déterminer laquelle de ces approches permet d'obtenir les meilleurs résultats. Dans l'application développée, le contexte représente un ensemble de peuplements partageant une fonction (production, réserve écologique, bloc expérimental ou terrain non productif) choisie par l'utilisateur. Le contexte est matérialisé par la TABLE DOMAINE (voir figure 6.4) qui regroupe tous les peuplements de tous les inventaires qui partagent la fonction choisie.

```
//initialisation de la table

requete.execute ("DROP TABLE TABLE_DOMAINE");

//la table domaine regroupe toutes les propriétés des
peuplements

requete.execute("CREATE TABLE TABLE_DOMAINE (peuplement
integer, age STRING, hauteur INTEGER, densité INTEGER,
Perturbation STRING, essence STRING, fonction STRING)");

//insertion des valeurs pour les peuplements de 1984 qui
possèdent la fonction choisie par l'utilisateur

requete.executeUpdate("INSERT INTO TABLE_DOMAINE SELECT
peuplement, age, hauteur, densité, Perturbation, essence,
fonction FROM Table_84 WHERE(Table_84.fonction LIKE '" +
fonction + "')");

//insertion des valeurs pour les peuplements de 1992 qui
possèdent la fonction choisie par l'utilisateur

requete.executeUpdate("INSERT INTO TABLE_DOMAINE SELECT
peuplement, age, hauteur, densité, Perturbation, essence,
fonction FROM Table_92 WHERE(Table_92.fonction LIKE '" +
fonction + "')");
```

Figure 6.4 : Constitution de la table domaine pour la détermination des poids

Les poids attribués aux différentes similarités sont ensuite déterminés à partir de la TABLE DOMAINE selon le principe probabiliste de la ressemblance ou de la variabilité (c. f. chap. 3.3). L'intégration du contexte dans le modèle de similarité sémantique permet d'attribuer plus d'importance aux propriétés particulières de ces peuplements lors de l'évaluation de la similarité globale, et ainsi de tenir compte de la distribution des propriétés des peuplements. La figure 6.5 montre les fréquences des valeurs de similarité obtenues à partir du principe de variabilité et de ressemblance pour le même contexte.

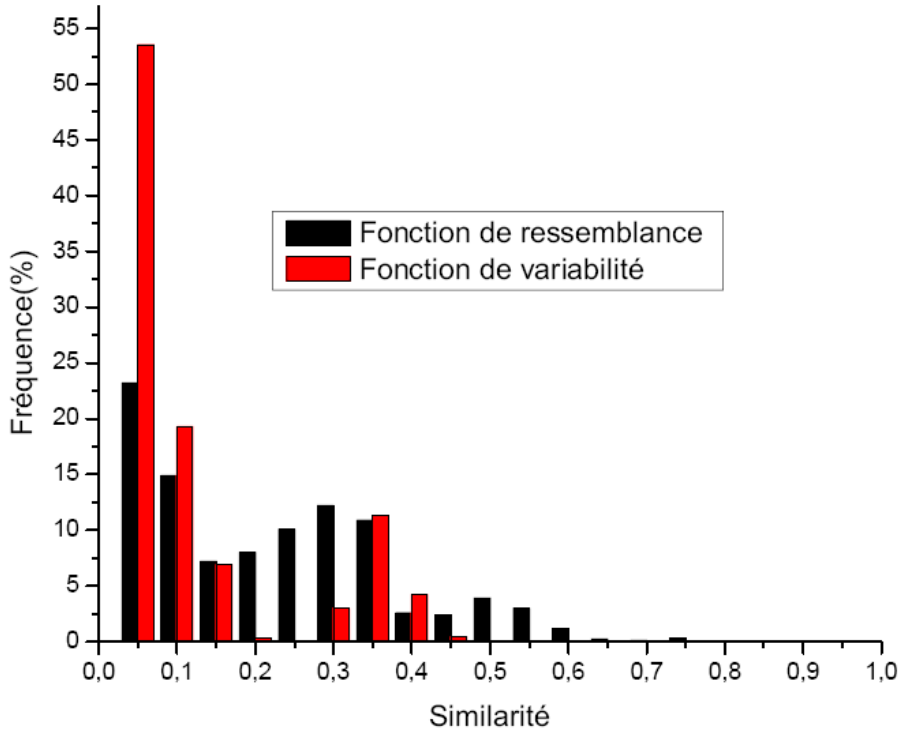

Figure 6.5 : Fréquence des valeurs de similarité pour le contexte production selon le principe de ressemblance et le principe de variabilité.

Les valeurs de similarité calculées selon le principe de variabilité sont en général significativement plus faibles que celles qui ont été calculées selon le principe de ressemblance, alors que les valeurs obtenues selon le principe de ressemblance sont distribuées de manière plus uniforme. Il est plus pertinent d'employer ce premier principe pour décrire la similarité entre des zones spatiales telles que les peuplements. En effet, le principe de ressemblance permet d'attribuer plus de poids aux propriétés qui sont partagées par les peuplements, alors que le principe de variabilité attribue plus d'importance aux propriétés qui distinguent les peuplements. De ce point de vue, la ressemblance est plus pertinente pour décrire le contexte forestier puisque les normes de gestion forestière se basent sur l'homogénéité des zones spatiales, c'est-à-dire qu'elles tendent à regrouper les peuplements qui partagent des propriétés semblables.

En appliquant le principe de ressemblance, les valeurs des poids obtenus à partir des différents contextes sont présentées dans le tableau 6.2.

Contexte	Poids
Production	Wa=0.6139, wf=0.2088, wp=0.1772
Réserve écologique	Wa=0.5829, wf=0.2231, wp=0.1930
Terrain non productif	wa=0.6176, wf=0.1911, wp=0.1911
Bloc expérimental	Wa=0.6315, wf=0.2105, wp=0.1578

Tableau 6.2 : Évaluation des poids avec le principe de ressemblance

Les valeurs de poids sont peu variables selon les contextes et, par conséquent, les résultats des fréquences de similarité obtenues selon différents contextes (figure 6.6) montrent que la variation de la distribution des valeurs de similarité est peu importante. La distribution suit une forme semblable, c'est-à-dire que la plupart des peuplements sont peu similaires (similarité comprise entre 0 et 0,10) ou moyennement similaires (similarité comprise entre 0,25 et 0.35). La faible variabilité des poids et, par conséquent, de la distribution des valeurs de similarité selon les différents contextes indique que les propriétés des instances (peuplements) sont peu variables par rapport à d'autres cas d'application, par exemple lorsque les entités sont tirés de l'ontologie WordNet et possèdent une grande variété de propriétés. Par exemple, l'essence *sapins* est nettement dominante (à toutes les époques) caractérisant plus de 50% des peuplements. De même, la hauteur peut prendre seulement quatre ou cinq valeurs, dépendamment des époques. Cette faible variabilité explique également que les valeurs de similarité obtenues à partir du principe de variabilité soient significativement plus faibles.

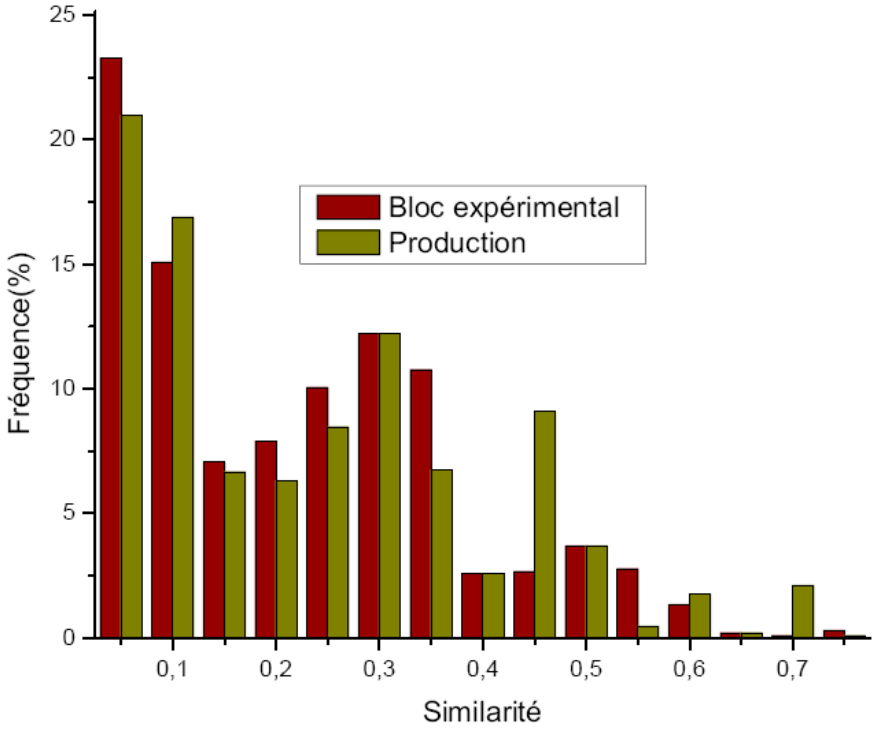

Figure 6.6 : Fréquence des valeurs de similarité selon le contexte *Bloc expérimental* et le contexte *Production.*

6. 4 Développement de l'application

L'objectif de cette section est d'expérimenter l'approche géosémantique proposée afin de montrer qu'elle permet de rétablir les liens sémantiques et géométriques entre les différents cubes de données géospatiales puis de répondre aux requêtes temporelles. L'application a été construite en utilisant le langage java et intègre les résultats de l'algorithme de rétablissement de liens sémantiques, les fonctionnalités du module Oracle Spatial pour l'approche géométrique, pour ensuite permettre de traiter selon l'approche géosémantique les requêtes spatio-temporelles. Finalement, l'application a été couplée avec l'outil SOLAP dans un prototype test pour mettre en valeur les possibilités de notre approche dans un processus d'analyse de l'évolution sémantique dans le domaine forestier.

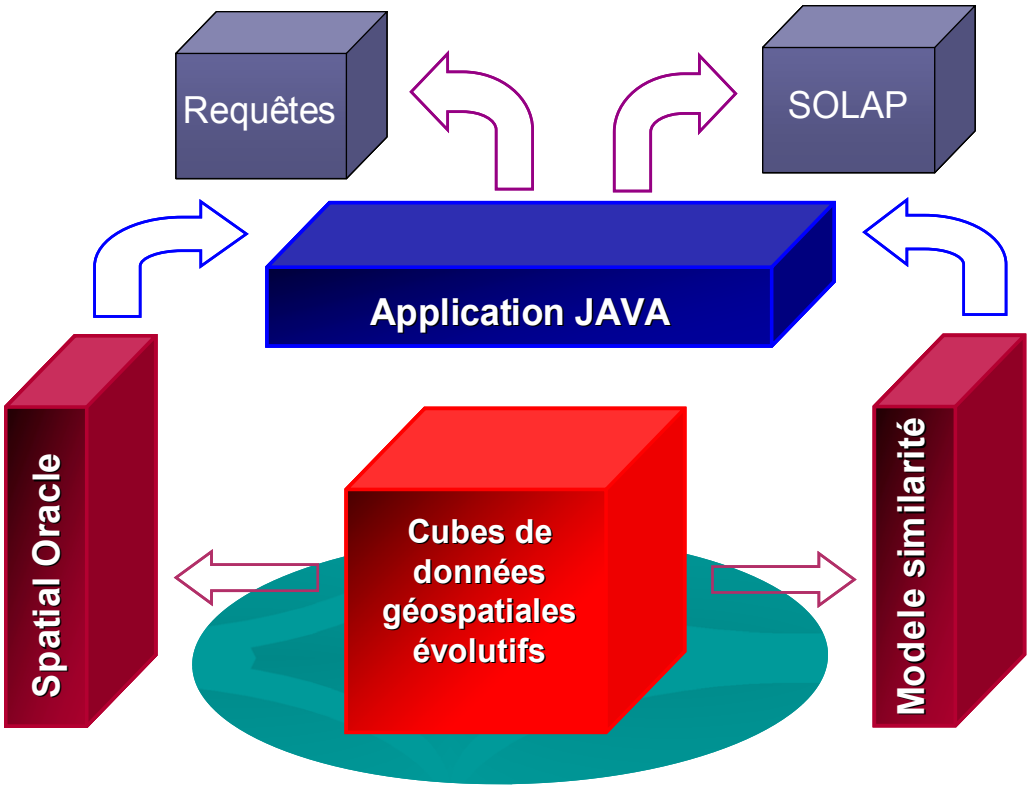

Figure 6.7: Schéma de l'architecture de l'application

6.4.1 Données de la forêt Montmorency

Les données proviennent des inventaires de la forêt Montmorency réalisés à différentes époques (1973, 1984, 1992 et 2002), chaque inventaire étant associé à un cube de données géospatiales. Les inventaires consistent à partitionner l'espace en zones présentant des propriétés homogènes en terme de densité, d'essences, de hauteur, etc., produisant un ensemble d'entités spatiales de base appelées peuplement. Les peuplements, associés à des polygones représentant leur géométrie, sont agrégés en entités spatiales de niveau supérieur et variables selon les époques (compartiment, polygone écologique, station forestière, etc.), formant la dimension spatiale *Découpage*. Les peuplements possèdent les propriétés décrites dans le tableau 6.3, lesquelles ont été déterminées lors des inventaires (à l'exception de la propriété *fonction*) et constituent les autres dimensions des cubes. Ces propriétés sont utilisées dans l'évaluation de la similarité sémantique. Les instances des

niveaux supérieurs (compartiments, polygones, stations, etc.) sont uniquement définis par les peuplements qui les composent.

Type de propriétés	Propriétés	Valeurs
Attributs	Densité Hauteur Âge Perturbation	[61,81]% [7,12] m [20,40] ans Coupe partielle
Parties	Essences	Bouleaux blancs et sapins : Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau blanc occupe plus de 50% de la surface terrière….
Fonction	Type de zone	Réserve écologique

Tableau 6. 3 : Exemple de propriétés d'une instance de peuplement forestier

En plus des propriétés recueillies pour chaque peuplement lors des inventaires, nous avons introduit la dimension *Fonction* qui permet de tenir compte des différentes fonctions d'un peuplement. La dimension Fonction possède un seul niveau hiérarchique et le schéma des instances est formé de quatre membres (figure 6.8).

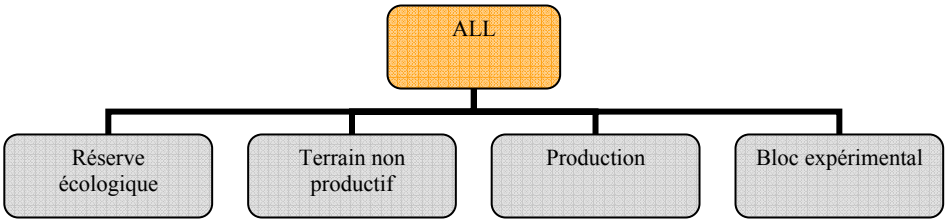

Figure 6.8 : Dimension *Fonction*

Les versions du cube de données forestières possèdent les mêmes dimensions (à l'exception de la dimension *Pente*, présente uniquement dans la version de 1992 et de 2001, n'ayant aucun analogue dans les versions précédentes) mais les membres des dimensions *Hauteur*, *Âge* et *Essence* ont évolué sémantiquement d'une version à l'autre. Par exemple, la définition de l'âge *prémature* est passée de [40,60 ans] à [45,60 ans] entre 1984 et 1992, et la majorité des textes décrivant les essences ont été modifiés. Les membres de la dimension *Découpage* sont définis (géométriquement et sémantiquement) de manière unique et

indépendante à chaque inventaire, par photo-interprétation et ne possèdent a priori aucun lien entre eux. Le tableau 6.4 présente un extrait des mesures qui furent utilisées dans l'application.

Id_Peuplements 1973	Surface (ha)	Densité moyenne (%)	Volume ligneux (m^3/ha)
1	3.1883	70.5	75.2
2	1.3297	50.5	85.3
3	1.6506	70.5	120.9
4	1.4512	70.5	0
5	0.3261	-	48.5
6	0.8974	70.5	12.2
7	0.9768	70.5	30.2
8	1.6633	50.5	90.4
9	0.7224	32.5	45.1

Id_Peuplements 1984	Surface (ha)	Densité moyenne (%)	volume ligneux (m^3/ha)
1818	2.2621	70.5	55.2
1817	3.4546	50.5	35.3
1802	8.9360	70.5	50.9
1816	2.3911	-	0
2358	1.4030	70.5	48.5
2395	0.2809	70.5	12.2
1812	2.0673	32.5	30.2
1803	5.9155	70.5	90.4
2357	1.1522	70.5	45.1

Id_Peuplements 1992	Surface (ha)	Densité moyenne (%)	Volume ligneux (m^3/ha)
3676	4,6448	-	0
3110	2,5434	90	82.9
3267	0,5568	-	0
3075	8,896	32.5	19.9
3194	0,9327	50.5	90.5
3268	0,1176	70.5	47.9
3254	0,24	70.5	47.9
3287	1,6033	70.5	100.9
3307	0,5745	70.5	47.9

Tableau 6.4: Extraits des surfaces, densité moyenne et estimation du volume ligneux pour les peuplements 1973, 1984 et 1992

L'évolution sémantique et géométrique fait en sorte que les peuplements de différentes époques ne sont pas liés entre eux et par conséquent, il est impossible de traiter les requêtes temporelles car les mesures du tableau 5.4 ne peuvent être comparées directement.

L'application de l'approche géosémantique permet de rétablir les liens sémantiques et géométriques entre les membres de la dimension spatiale et de traiter les requêtes temporelles portant sur plusieurs cubes de données géospatiales.

6.4.2 Implémentation et application de l'approche

Dans un premier temps, les résultats de l'application permettent de montrer les liens sémantiques établis entre les peuplements. L'interface permet à l'utilisateur de diriger le processus de rétablissement de liens sémantiques en spécifiant le contexte d'évaluation de la similarité au moyen du choix d'une fonction (production, réserve écologique, bloc expérimental, etc.) et offre également la possibilité de spécifier manuellement la valeur des poids (fig. 6.9), dans le cas où, par exemple, une seule composante de la similarité serait intéressante. Cette option peut être utile par exemple si l'utilisateur veut rétablir les liens en tenant compte uniquement des parties des instances (dans ce cas, des essences), ou encore s'il veut connaître l'évolution sémantique en considérant uniquement un type de propriété à la fois.

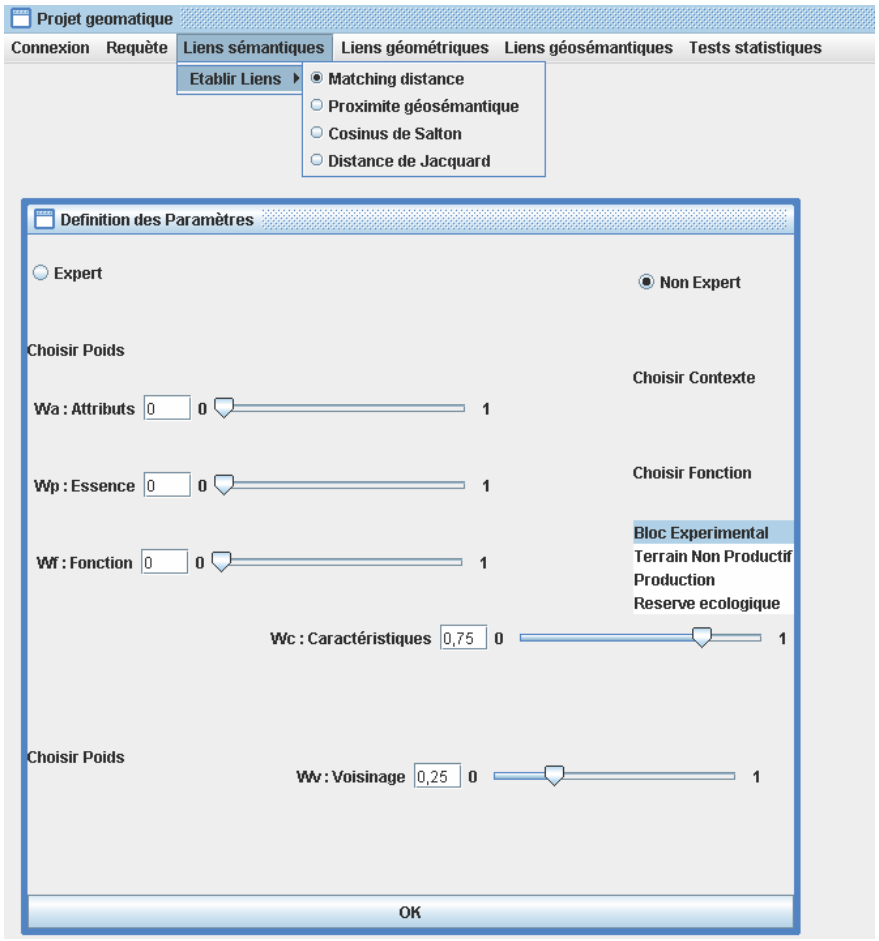

Figure 6.9 : Définition des paramètres pour le rétablissement des liens sémantiques

Chaque matrice de correspondances sémantiques créée lors de ce processus lie les membres appartenant à deux niveaux de la dimension spatiale *Découpage* de deux époques différentes. Une fois les liens sémantiques établis, l'utilisateur peut choisir de visualiser une des quinze matrices de correspondances sémantiques résultantes (fig. 6.10).

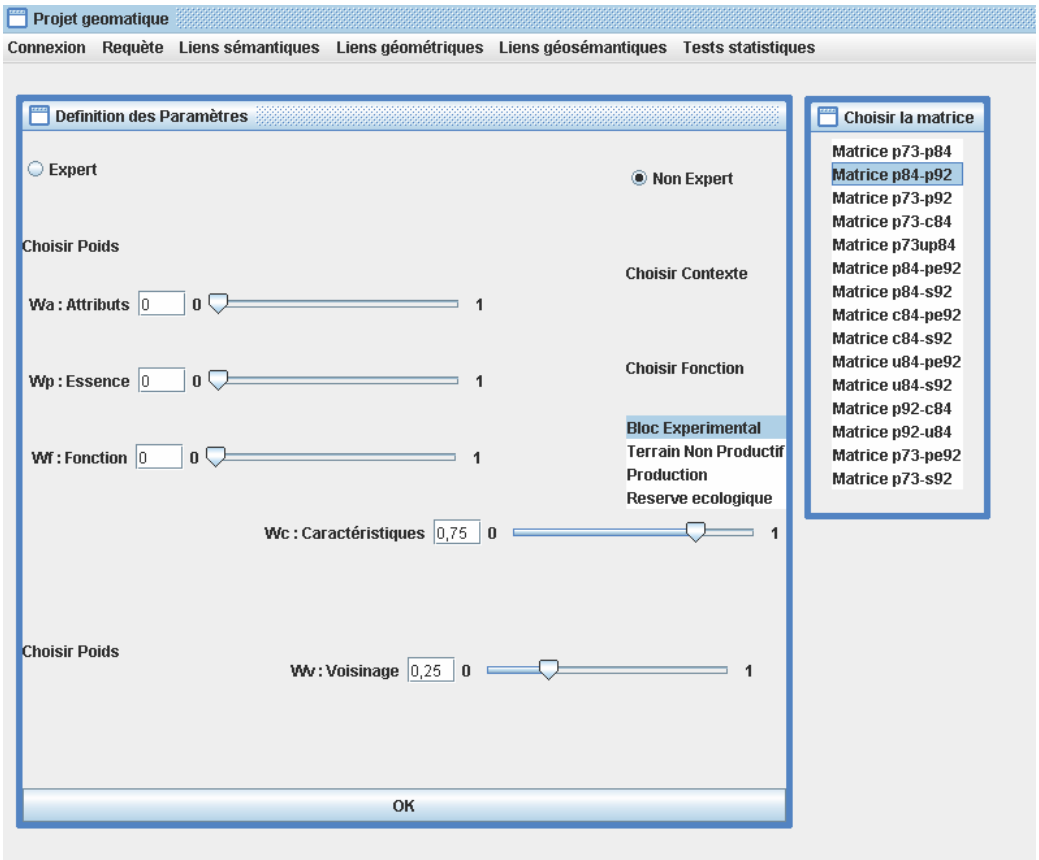

Figure. 6.10 : Matrices de correspondances sémantiques créées lors du processus de rétablissement de liens sémantiques

La figure 6.11 montre un exemple de matrice de correspondances sémantiques reliant les niveaux peuplement forestier (1984) et peuplement écoforestier (1992). L'annexe B montre des extraits des matrices de correspondances sémantiques qui furent obtenues selon différents niveaux et différents contextes. Dans l'annexe A (tableaux 1A à 3A), les propriétés de chaque peuplement pour lesquels la similarité a été évaluée sont listées. Les codes employés pour représenter les propriétés (essences, âge, hauteur, densité) sont indiqués dans les tableaux 4A à 7A. Aucune sélection n'est effectuée à partir de la valeur des liens sémantiques, c'est-à-dire que toutes les valeurs contenues dans les matrices de correspondances sémantiques seront considérées pour répondre aux requêtes temporelles.

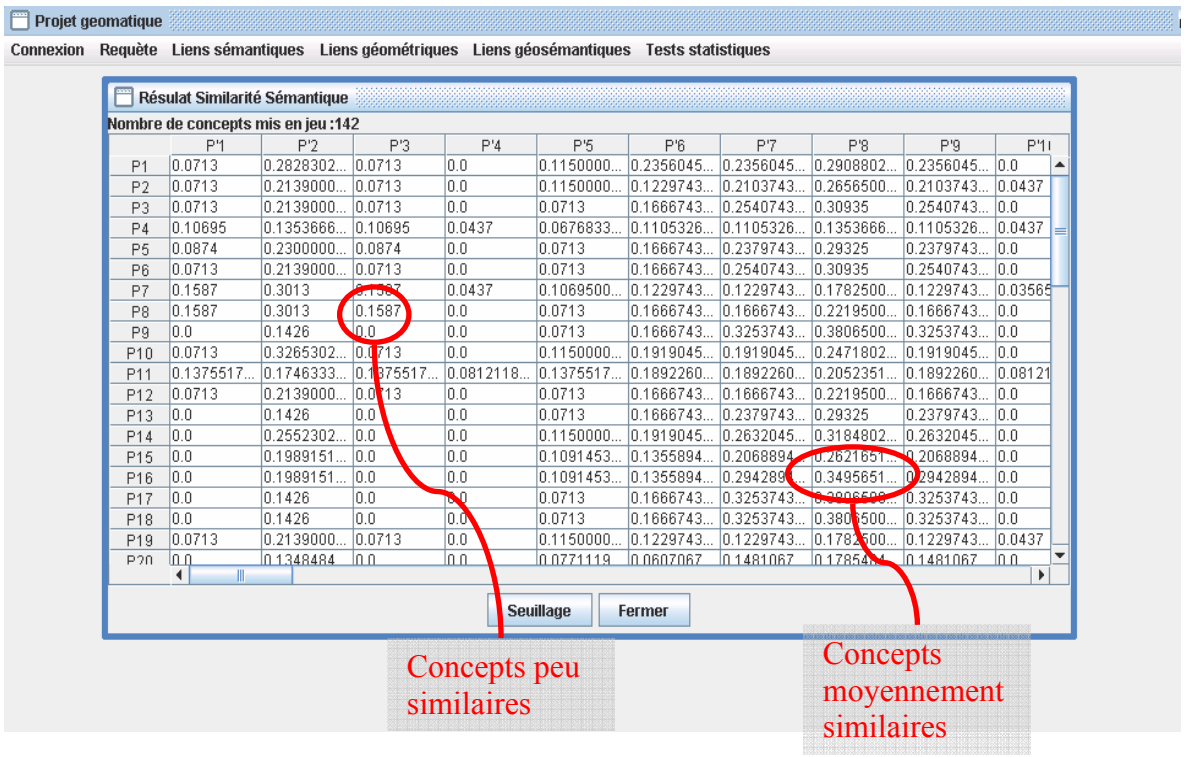

Figure 6.11 : Visualisation d'une matrice de correspondances sémantiques

Les matrices de correspondances sémantiques sont utilisées pour constituer les tables de stockage de liens sémantiques qui seront nécessaires pour traiter les requêtes temporelles. Les tables de stockage des liens sémantiques consignent tous les couples d'instances appartenant à deux niveaux de deux inventaires différents qui présentent une similarité non nulle. Elles sont constituées en parcourant séquentiellement les cellules des matrices de correspondances sémantiques. Le tableau 6.5 montre un exemple d'un extrait d'une des tables de stockage qui fut constituée dans le cadre de notre approche et qui indique les peuplements de 1992 qui présentent une similarité non nulle avec un peuplement de l'inventaire 1984.

Peuplements 1984	Peuplements 1992
2352	2847
2352	3176
2342	3440
1813	3110
1813	3075
…	…

Tableau 6.5 : Exemple d'un extrait d'une table de stockage des liens sémantiques

De la même manière, les matrices de correspondances géométriques créées suite à la phase d'indexation sont utilisées pour constituer les tables de stockage de liens géométriques. Les tables de stockage des liens géométriques consignent tous les couples d'instances appartenant à deux niveaux de deux inventaires différents qui présentent une géométrie non disjointe, c'est-à-dire un taux d'inclusion différent de zéro. Elles sont constituées en parcourant séquentiellement les cellules des matrices de correspondances géométriques. Le tableau 6.6 montre un exemple d'un extrait d'une des tables de stockage de liens géométriques qui fut constituée dans le cadre de notre approche et qui indique les peuplements de 1992 qui présentent un taux d'inclusion non nul avec un peuplement de l'inventaire 1984.

Peuplements 1984	Peuplements 1992
1818	3419
1818	2976
1818	3440
1817	3286
1817	2960
1817	2779
1817	3254
1802	3440
1802	2779
…	…

Tableau 6.6 : Exemple d'un extrait d'une table de stockage des liens géométriques

Finalement, les tables de stockage des liens géosémantiques stockent tous les couples d'instances qui sont présents à la fois dans la table de stockage des liens sémantique et la table de stockage des liens géométriques. Elles sont constituées en parcourant premièrement les tables de stockage des liens géométriques et en identifiant parmi les couples de cette table ceux qui sont également présents dans la table de stockage des liens sémantiques. Les tables de stockage des liens géosémantiques sont utilisées lors du traitement des requêtes temporelles. Lorsque l'utilisateur spécifie une requête pour un peuplement d'un inventaire donné, elles permettent de repérer les peuplements correspondants (sémantiquement et géométriquement) dans un autre inventaire touché par la requête.

6.4.3 Prototype test SOLAP

Nous avons réalisé un prototype SOLAP, l'extension spatiale de l'outil OLAP, à partir des résultats de notre application, afin de pouvoir visualiser cartographiquement les liens sémantiques établis. La figure 6.12 montre le schéma en étoile de la structure multidimensionnelle réalisée. Les faits sont les mesures de similarité sémantique évaluées par rapport aux différentes dimensions que sont les couples de peuplements, les essences, la perturbation et les poids des attributs, parties, fonctions, du voisinage et de la similarité lexicale. Les membres des dimensions poids_attributs, poids_parties, poids_fonctions, poids_voisinage et poids_lexical sont des catégories de valeurs pour ces poids (*très forte similarité*, *forte similarité*, *moyenne similarité* et *pas de similarité*). Les membres de la dimension *couple_de _peuplements* sont des paires de peuplements appartenant à deux inventaires différents. Les membres des dimensions Essences et Perturbation sont les mêmes que ceux qui ont été définis dans les inventaires.

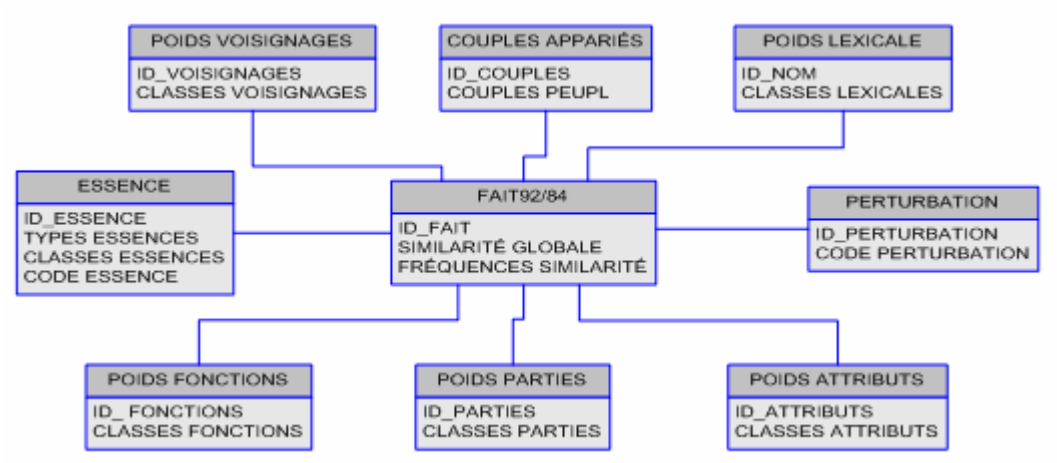

Figure 6.12 : Schéma en étoile du cube test de similarité sémantique

Les résultats montrent que la mesure de similarité sémantique présente un potentiel intéressant pour analyser les données géospatiales. En effet, la similarité, lorsque visualisée cartographiquement, permet d'indiquer la valeur des liens sémantiques entre les peuplements ainsi que le degré d'évolution sémantique des zones de la forêt. Ainsi, l'utilisateur peut rapidement apercevoir quelles régions de la forêt ont été plus affectées par des changements de différentes natures, ou encore celles qui sont restées stables. Les

différents poids qui déterminent la similarité sémantique globale prennent alors beaucoup de sens puisqu'ils permettent à l'utilisateur de visualiser l'évolution selon différentes dimensions, soit celles des attributs (âge, hauteur, densité…), des parties (les essences) ou des fonctions des peuplements. Ce type d'analyse offre une vision globale pouvant agir à titre de première analyse avant de permettre à l'utilisateur de mener des analyses plus précises, facilitant ainsi la prise de décision en indiquant rapidement les zones susceptibles de présenter plus d'intérêt. Imaginons en effet qu'un utilisateur veuille identifier les zones de la forêt où les essences ont, par exemple, faiblement évolué. Il devra effectuer un très grand nombre de requête, par hasard, avant d'identifier ces zones alors qu'en visualisant directement les zones où la similarité des parties (essences) se situe, par exemple, entre 0,25 et 0,35, il peut rapidement cerner les peuplements concernés. La figure 6.13 montre une requête où l'on peut visualiser la distribution de la similarité entre le peuplement 2 de l'inventaire de 1984 et les peuplements 1 à 15 de l'inventaire de 1992, pour des valeurs de poids choisies par l'utilisateur. Par exemple, les peuplements caractérisés par une similarité comprise dans l'intervalle [0.387, 0.4837] sont représentés en rouge, les différentes couleurs permettant de visualiser très rapidement les zones plus ou moins similaires.

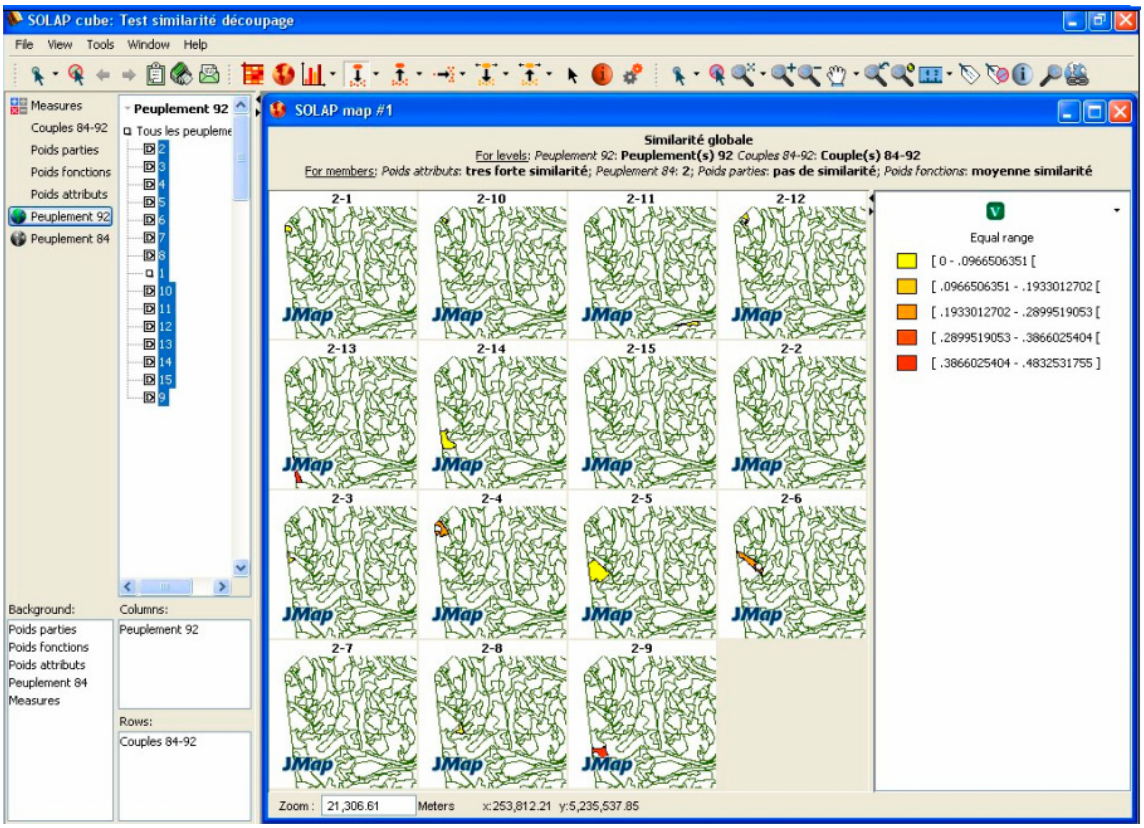

Figure 6.13 : Exemple de requête sur la distribution spatiale des similarités.

Un autre exemple de requête possible est d'identifier les couples de peuplements, entre 1984 et 1992) dont la similarité se situe entre une intervalle donné, dans cet exemple entre 0.48 et 0.50 (figure 6.14).

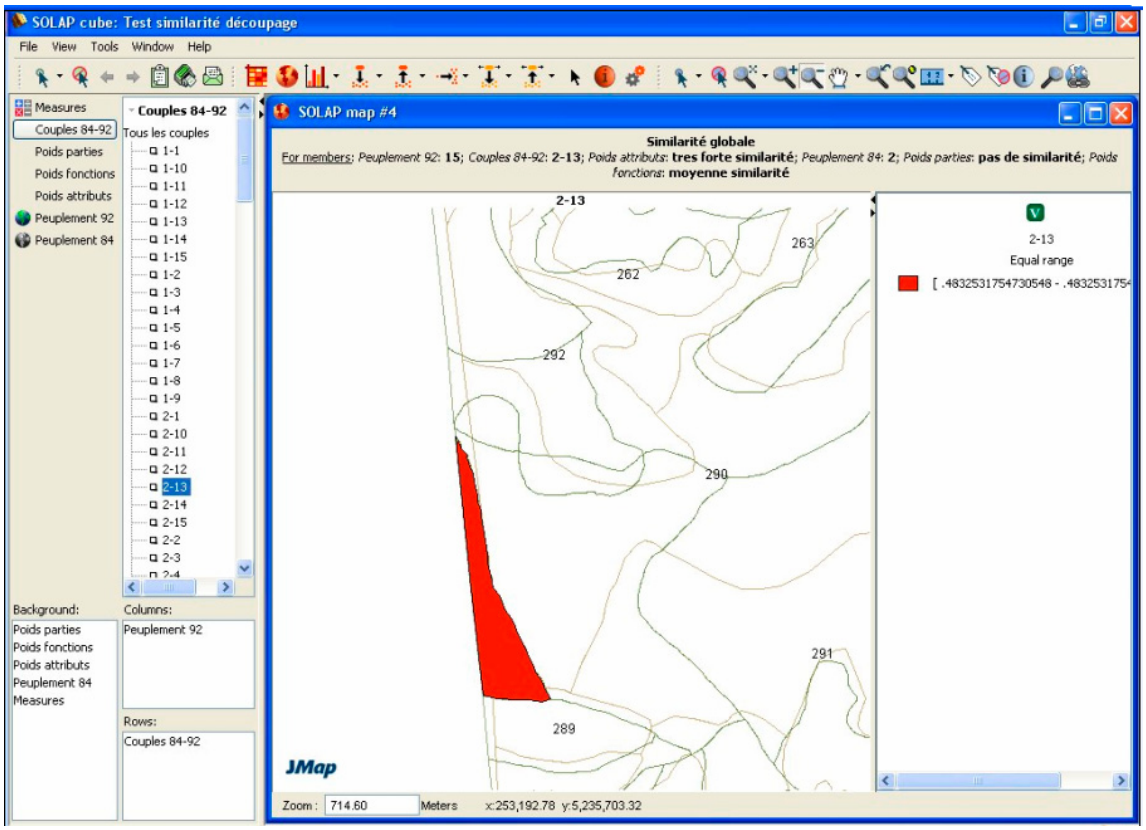

Figure 6.14 : Exemple de requête sur les peuplements se situant dans un intervalle de similarité donné

La mesure de similarité couplée à l'outil SOLAP peut donc être un moyen efficace et rapide de visualiser les liens sémantiques et l'évolution sémantique des peuplements forestiers.

6.4.4 Requêtes spatio-temporelles

Le traitement des requêtes spatiotemporelles avec l'approche géosémantique constitue l'objectif principal de l'application développée. Dans le chapitre précédent, nous avons présenté comment la combinaison des approches sémantiques et géométriques avec la

méthode de transformation permet de transformer les mesures du cube pour répondre à des requêtes spatio-temporelles sur des membres de la structure spatiale qui, jusqu'alors, demeuraient indépendants les uns des autres. Nous présentons ici concrètement comment l'algorithme développé effectue le processus de traitement des requêtes. Lors de la requête, l'utilisateur spécifie les paramètres suivants :

- Une ou plusieurs entités cartographiques (membres des dimensions spatiales *Découpage*) appartenant à une version *VD* (1973, 1984, 1992 ou 2002);
- une mesure (par exemple, la surface ou le volume ligneux);
- un intervalle de temps $[t_i, t_f]$ pour lequel l'utilisateur souhaite connaître les données;
- un domaine de valeurs pour un ou plusieurs attributs descriptifs (par exemple essence = épinette et sapins);
- une fonction d'agrégation des données dans le temps (par exemple, la moyenne ou l'évolution des données);
- une version de représentation des données Vr, telle que tous les résultats de la requête seront transposés dans cette version.
- Le type de requête : *avec liens sémantiques*, où seule la matrice de correspondances sémantiques est employée pour transformer les mesures ; *avec liens géométriques*, où seule la matrice de correspondances géométriques est employée pour transformer les mesures ; *avec liens géosémantiques*, où les mesures sont transformées selon la matrice de correspondances géosémantique (c.f. section 5.5.2).

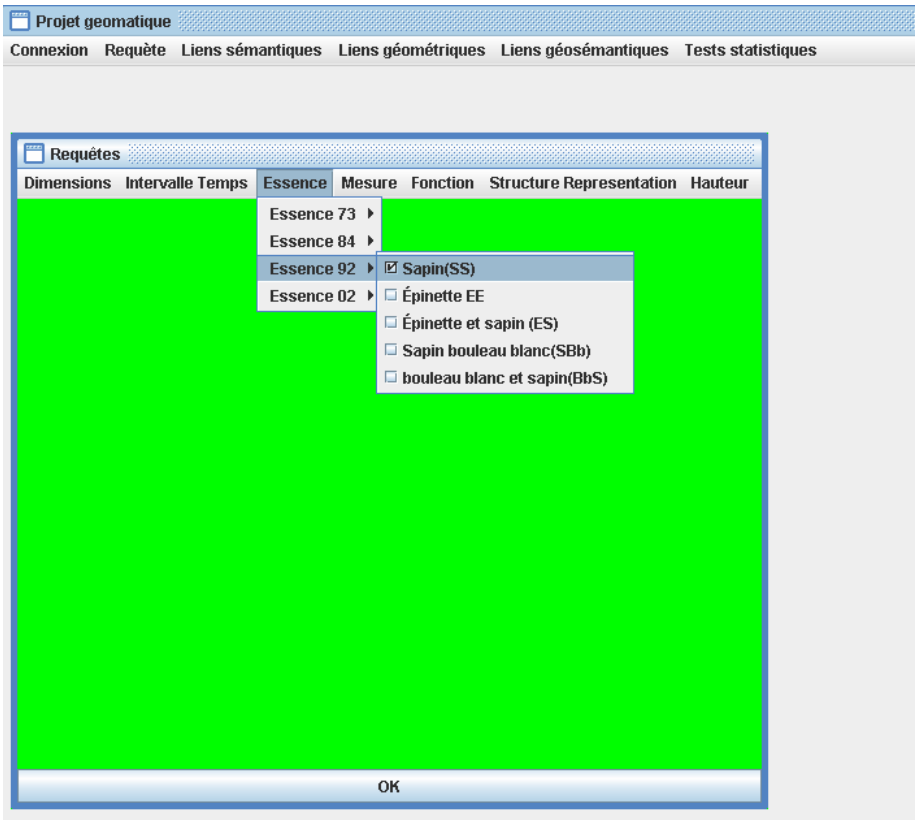

Figure 6.15 : Exemple de spécification d'un attribut descriptif *Essence* lors d'une requête

Les requêtes peuvent porter sur les entités cartographiques d'un inventaire que l'utilisateur est libre de choisir. Ainsi, plutôt que de créer une seule représentation cartographique accessible, l'utilisateur peut effectuer une analyse à partir d'un découpage forestier effectué à une époque de son choix. Ceci offre plus de possibilités pour l'utilisateur qui peut choisir une représentation avec laquelle il est déjà familier plutôt que de se voir imposer un découpage qui ne correspond pas à la réalité forestière. Même si en général le découpage le plus récent est le plus adapté, il est possible qu'un utilisateur préfère une représentation ultérieure.

L'algorithme de traitement des requêtes selon l'approche géosémantique effectue les étapes suivantes (figure 6.16) :

- Identification de l'ensemble V des *n* versions correspondantes à la requête : par exemple si la requête porte sur l'intervalle [t_i= 1984, t_f = 2002], l'ensemble des versions est donné par V= {1984, 1992 et 2002};

- Séparation de la requête en *n* requêtes partielles, soit une requête pour chaque version de la structure;

- Traitement des requêtes partielles, lequel comprend trois phases :

  1) Extraction : les mesures correspondant au peuplement de la requête sont extraites du cube de données et sont représentées sous forme matricielle, ou chaque élément de la matrice correspond à une cellule du cube (c.f. section 4.5).

  2) Identification des peuplements correspondants au peuplement de la requête par la table de stockage des liens géosémantiques et extraction des mesures du cube liées à ces peuplements.

  3) Transformation des mesures extraites en 1) et 2) : deux cas possibles se présentent : (i) si la matrice du cube est extraite de la version Vi correspondant à la structure de représentation Vr, les mesures se trouvent déjà dans la structure de représentation choisie et ne nécessitent pas de transformation matricielle ; (ii) si la matrice du cube est extraite d'une version Vi différente de la structure de représentation Vr, la matrice des mesures doit être transformée dans la version choisie Vr en utilisant les matrices de correspondances. Si le type de requête choisi est *avec liens sémantiques*, seule la matrice de correspondances sémantiques est utilisée comme matrice de transformation. Si le type de requête choisi est avec liens géosémantiques, la matrice de liens géosémantique est utilisée comme matrice de transformation.

- Une fois que les résultats des requêtes partielles, qui prennent la forme de matrice, sont déterminés, les résultats peuvent être combinés avec une fonction choisie par l'utilisateur (moyenne, évolution).

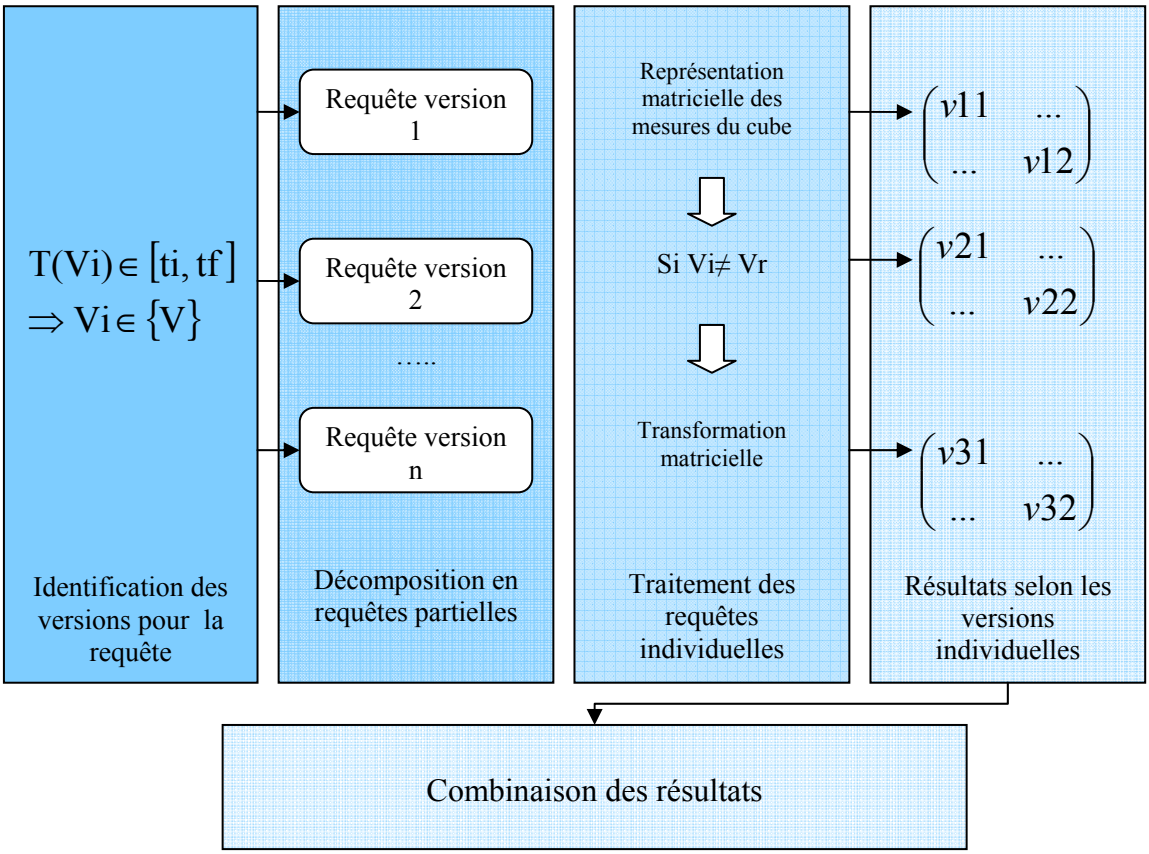

Figure 6.16 : Processus de traitement des requêtes temporelles

Le type de requête *avec liens sémantiques* s'applique pour des données qui seraient uniquement descriptives ou encore dans un contexte ou la géométrie associée aux instances est demeurée inchangée. La figure 5.17 montre un exemple de requête *avec liens sémantiques* où l'utilisateur veut connaître la moyenne du volume ligneux pour l'essence *sapins* des peuplements 1818, 1817 et 1802 de l'inventaire de 1984 entre 1984 et 1992. L'utilisateur a spécifié que les résultats devaient être représentés dans la structure de 1984. Pour le rétablissement des liens sémantiques, l'utilisateur a considéré le contexte de production. La matrice de correspondances sémantique utilisée est présentée à la table 1 de l'annexe B, tableau 5B.

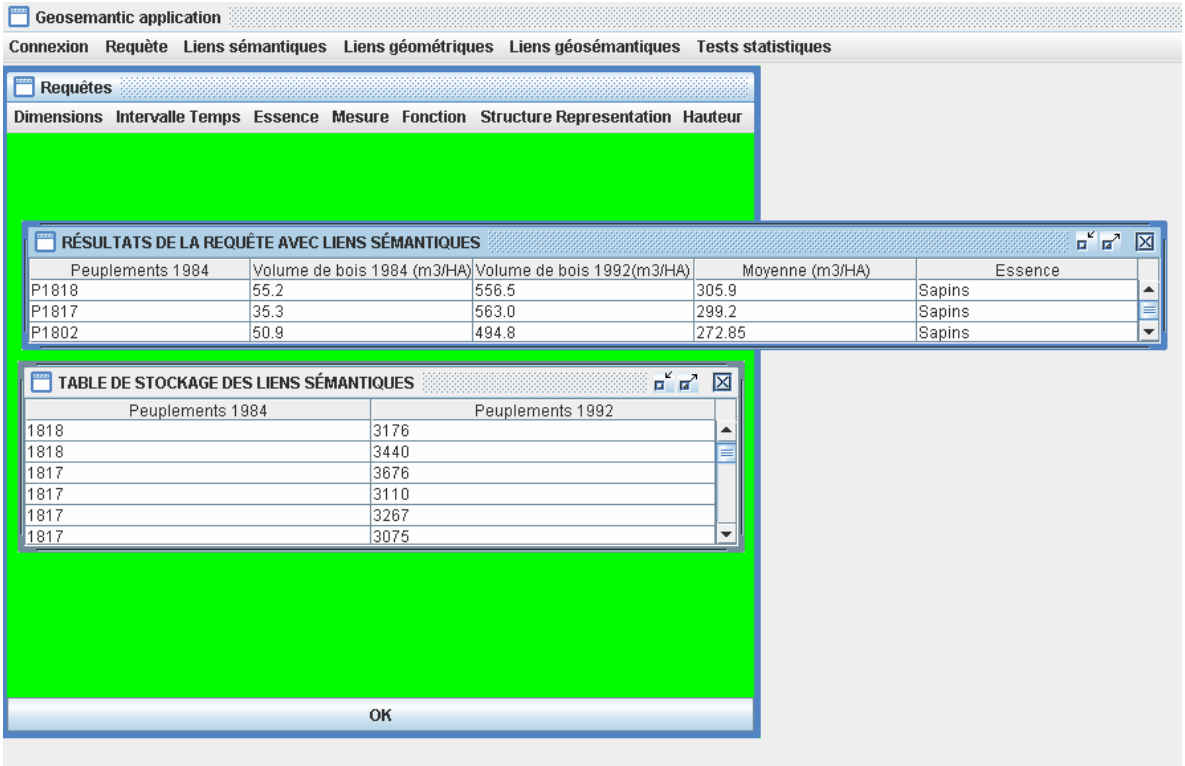

Figure 6.17 : Résultat d'une requête *avec liens sémantiques*

Peuplements	Volume de bois 1984	Volume de bois 1992	Moyenne	Essence
1818	55.2	556.5	305.9	Sapins
1817	35.3	563.0	299.2	Sapins
1802	50.9	494.8	272.85	Sapins

La première table de la figure 6.17 indique les résultats obtenus pour le volume de bois de chaque peuplement, en 1984, en 1992 puis la moyenne de ces résultats. La seconde table montre les peuplements de 1992 qui sont associés sémantiquement avec les peuplements 1818, 1817 et 1802 (issus de l'inventaire de 1984) de la requête. Les tables de stockage des liens sont obtenues en identifiant dans la matrice de correspondances sémantiques les couples de peuplements qui sont liés par une valeur de similarité non nulle. Puisque le traitement des requêtes s'appuie uniquement sur les liens sémantiques, le résultat obtenu tient compte de tous les peuplements de l'inventaire qui partagent un degré quelconque de similarité avec les peuplements 1818, 1817 et 1802, sans considération pour les relations géométriques qui les caractérisent. Par conséquent, le volume de bois calculé en 1992 est

très élevé par rapport à celui de 1984 et ne représente pas le volume réel pouvant être associé aux peuplements de la requête. Par contre, compte tenu que la version de représentation choisie est celle de 1984, les résultats du champ *Volume de bois 1984* sont justes puisqu'ils ont uniquement été extraits du cube de données de 1984 (sans être transformés). Ces résultats montrent que dans notre contexte, il est nécessaire de tenir compte des relations géométriques entre les peuplements pour donner un résultat valide. Cependant, il est concevable que dans certaines situations particulières, une requête traitée uniquement avec les liens sémantiques puisse donner des résultats valides. Les conditions nécessaires qui permettent de considérer uniquement les liens sémantiques sont les suivantes :

- La géométrie des peuplements est demeurée inchangée ;
- Les correspondances entre les peuplements de différents inventaires sont sans équivoques, c'est-à-dire que les correspondances sémantiques sont établies sur la base de propriétés exclusives qui assurent l'unicité d'un phénomène. Ainsi, si deux instances sont identifiés comme similaires, par exemple, il n'existe pas d'autres concept qui puisse leur être identique.
- Les correspondances sont de type simple (1: n ou n : 1) et non multiple (n : m), sans quoi les mesures associées à un peuplement d'une époque donnée peuvent être attribuées à plusieurs peuplements d'une autre époque (sans qu'il n'y ait division).

En ce qui concernent les données forestières utilisées, les résultats de l'évaluation de la similarité sémantique montrent que ces cas ne sont pas respectées, puisque (1) les géométrie ont toutes été modifiées; (2) les correspondances sémantiques n'indiquent pas un correspondance réelle entre deux zones: deux peuplements peuvent présenter des propriétés identiques mais être situés dans des régions lointaines du territoire considérés, il serait alors faux de les faire coïncider ; (3) les correspondances sont de type multiples. Par conséquent, dans le contexte forestier, même si la géométrie est demeurée inchangée, il est nécessaire de tenir compte des correspondances géométriques afin que les résultats de la requête soit cohérent avec la zone questionnée.

Le type de requête *avec liens géométriques* peut s'appliquer pour des données où la sémantique est demeurée inchangée, c'est-à-dire quand la similarité entre les peuplements de chaque époque vaut 1. La figure 6.18 montre les résultats de la requête précédente *avec liens géométriques*.

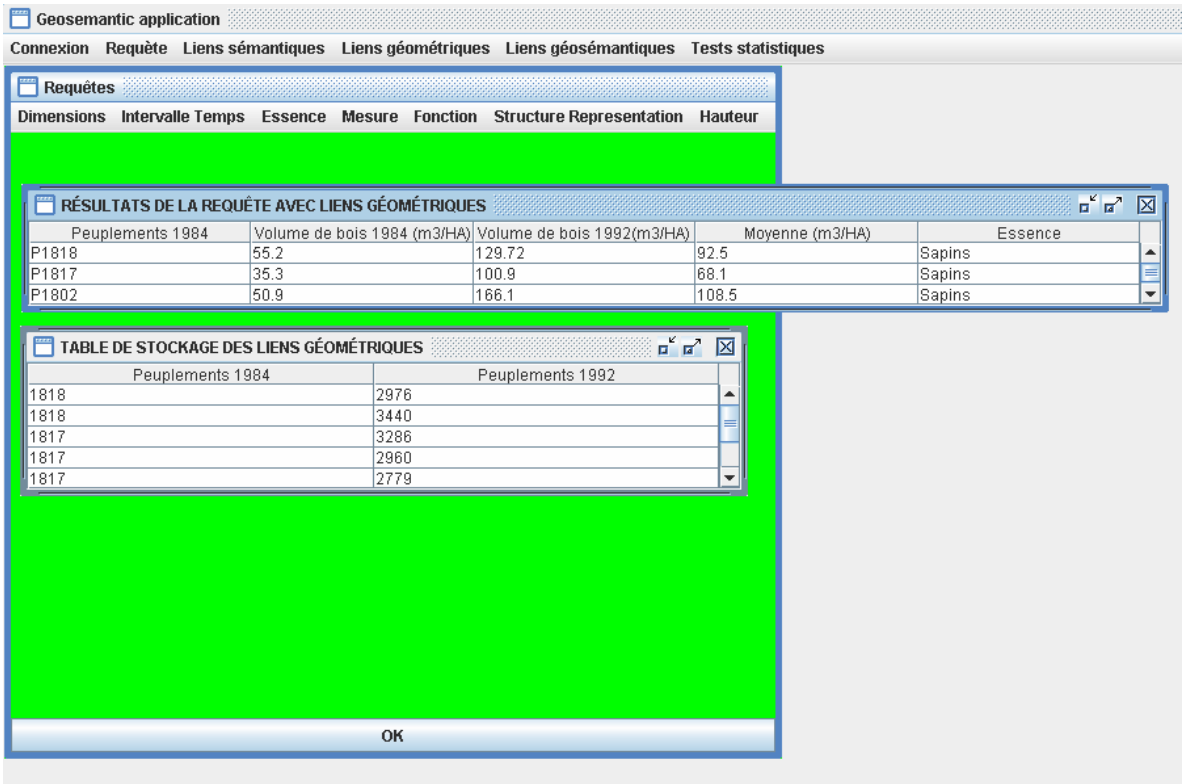

Figure 6.18 : Résultat d'une requête *avec liens géométriques*

Peuplements 1984	Volume de bois 1984	Volume de bois 1992	Moyenne	Essences
1818	55.2	129.7	92.5	Sapins
1817	35.3	100.9	68.1	Sapins
1802	50.9	166.1	108.5	Sapins

Comme dans le cas de la requête *avec liens sémantiques*, le résultat obtenu tient compte uniquement des correspondances géométriques, c'est-à-dire du taux d'inclusion des peuplements 1818, 1817, 1802 avec les peuplements de 1992. La table de stockage des liens géométriques permet d'identifier l'évolution géométrique en indiquant les peuplements de 1992 qui présentent une géométrie non disjointe de la géométrie des

peuplements de la requête. Par exemple, la surface du peuplement 1818 de l'inventaire de 1984 est non disjointe de la surface des peuplements 2976 et 3440 de l'inventaire de 1992, c'est-à-dire que la zone du peuplement 1818 a été divisée en deux zones. Le taux d'inclusion des peuplements 2976 et 3440 avec le peuplement 1818 donne le pourcentage du volume de bois de ces peuplements qui est considéré pour donner le résultat du volume de bois en 1992. Cependant, ces résultats sont calculés sans tenir compte de l'évolution sémantique entre les peuplements, par exemple du fait que la définition de l'essence Sapins a été modifiée entre 1984 et 1992 et que, par conséquent, les deux peuplements ne comportent pas le même pourcentage de sapins par rapport à leur surface.

Les résultats des types de requête *avec liens sémantiques* et *avec liens géométrique* montrent que ces approches ne peuvent s'employer que dans des cas particuliers et que dans le contexte forestier, il est nécessaire de tenir compte simultanément de l'évolution sémantique et géométrique pour obtenir des résultats plus près de la réalité. La figure 6.19 montre les résultats obtenus pour les peuplements 1818, 1817 et 1802 cette fois en considérant les liens sémantiques et géométriques. La table de stockage des liens géosémantiques indique les peuplements de 1992 qui possèdent à la fois une valeur de similarité non nulle et une géométrie non disjointe par rapport aux peuplements de la requête.

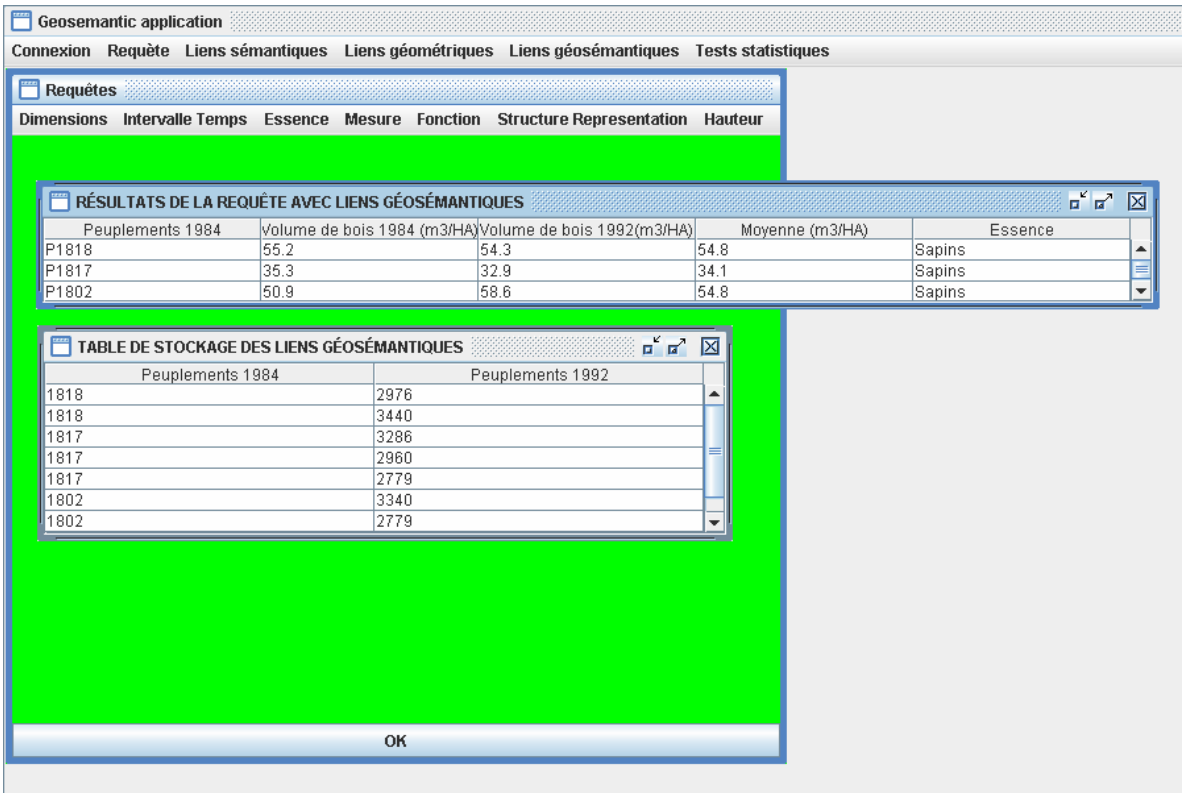

Figure 6.19 : Résultat de la requête *avec liens géosémantiques*

Peuplements 1984	Volume de bois 1984	Volume de bois 1992	Moyenne	Essences
1818	55.2	54.3	54.8	Sapins
1817	35.3	32.9	34.1	Sapins
1802	50.9	58.6	54.8	Sapins

6. 5 Analyse des résultats et discussion

Dans la section précédente, nous avons montré que l'approche géosémantique permet de traiter les requêtes spatio-temporelles sur des données provenant de cubes de données géospatiales de différentes époques. Ces résultats sont encourageants en soi puisque a priori, les liens entre les membres de la dimension spatiale des différents inventaires ne sont pas conservés suite à l'évolution de la structure multidimensionnelle. Par exemple, il était impossible de connaître l'évolution du volume ligneux pour le peuplement 2976 de l'inventaire de 1992 puisqu'il ne pouvait être identifié à aucun autre peuplement dans les inventaires précédents. L'approche géosémantique permet donc de donner des résultats

dans une situation où ils étaient impossibles à obtenir. De plus, les résultats tendent à se rapprocher des valeurs réelles puisqu'ils tiennent compte de l'évolution sémantique des propriétés des peuplements (évolution de la dimension âge, hauteur, essences, etc.) et de leur évolution géométrique. Pour évaluer davantage la qualité des résultats, nous avons choisi de procéder à une analyse comparative entre les résultats obtenus en utilisant le modèle de similarité Matching Distance et ceux qui furent obtenus avec le modèle de similarité sémantique redéfini. Comme il en fut discuté dans le chapitre 3, le modèle MD ne considère la similarité sémantique qu'au niveau des concepts, mais pas entre les propriétés des concepts. Par conséquent, nous supposons que les résultats des requêtes obtenus avec le modèle redéfini seront plus précis puisqu'ils considèrent la similarité entre les propriétés. En confrontant les résultats des requêtes qui s'appuient sur les deux modèles, nous visons à évaluer l'apport du modèle redéfini dans le traitement des requêtes spatio-temporelles. La démarche élaborée pour l'analyse des résultats est illustrée à la figure 6.20. Dans la première phase de ce processus de validation, nous avons obtenus les matrices de correspondances sémantiques en appliquant le modèle de similarité sémantique Matching Distance (MD) et le modèle de similarité sémantique redéfini (MDR), tel qu'elles ont été présentées dans la section 6.2. Les exemples de matrices de correspondances sémantiques obtenues pour le modèle MD et le modèle redéfini sont montrés en annexe B. Lors de la seconde phase, nous avons effectué une itération d'une requête sur plusieurs n= 135 peuplements en considérant uniquement les liens sémantiques, la requête demandant le volume ligneux pour le peuplement i (i=1,2,3,….n) pour l'essence sapin entre 1984 et 1992. Les deux ensembles de résultats obtenus pour cette série de requêtes furent comparés en évaluant le taux d'écart entre les valeurs fournies par le modèle MD et les valeurs fournies par le modèle de similarité sémantique redéfini. Puis, afin de tenir compte de la réalité du contexte forestier, nous avons intégré dans le processus de traitement des requêtes les matrices de correspondances géométriques obtenues avec la méthode d'indexation *QuadTree* de spatial oracle. Ceci nous a permis d'obtenir une seconde série de résultats aux mêmes requêtes en tenant compte cette fois des liens sémantiques et géométriques, c'est-à-dire selon l'approche géosémantique intégrée. Pour cette seconde série de résultats, le taux d'écart fut également évalué et comparé au taux d'écart obtenu avec les premiers résultats.

Figure 6.20 : Processus d'analyse des résultats des requêtes spatio-temporelles

Le tableau 6.6 montre les résultats obtenus, d'une part la moyenne des résultats aux séries de requêtes pour le volume ligneux obtenu avec chacun des modèles et d'autre part, l'écart moyen entre les résultats des deux modèles.

	Avec modèle MD	**Avec modèle MDR**	
	Moyenne de la mesure du volume (m3/ha)	Moyenne de la mesure du volume (m3/ha)	**Écart moyen (m3/ha)**
Requêtes avec liens sémantiques	109,5	241,8	132,4
Requêtes avec liens géosémantiques	4,1	19,9	15,1

Tableau 6.6 : Résultats de l'analyse comparative pour l'approche géosémantique intégrant le modèle MD et le modèle redéfini

- **Requête avec liens sémantiques** : le résultat du volume est considérablement plus faible avec le modèle MD, parce que ce dernier ne considère pas la similarité à un niveau suffisamment précis (c'est-à-dire entre les propriétés des instances) par rapport au modèle redéfini. Le modèle MD ne permet pas, en réalité, de tenir compte de l'évolution sémantique des membres du schéma des instances de la dimension au plus fin niveau. Dans les deux cas (avec modèle MD et modèle MDR), les résultats sont élevés par rapport aux données pour chaque peuplement (par exemple, selon le tableau 5.4, la moyenne des volumes pour les peuplements de 1992 est de 47,8) car la requête avec liens sémantiques seulement additionne une partie des volumes ligneux de tous les peuplements qui possèdent un degré de similarité non nul avec le peuplement de la requête (peuplement *i*). Par conséquent, dans le contexte forestier, il est nécessaire de tenir compte des liens géométriques entre les membres des dimensions spatiales. Toutefois, ce résultat montre que l'écart moyen entre les résultats des requêtes obtenues avec le modèle MD et le modèle redéfini est important et que par conséquent l'apport de la redéfinition du modèle ne peut être négligé.

- **Requête avec liens géosémantiques** : la moyenne du volume pour les requêtes avec liens géosémantiques est plus faible que la moyenne obtenue avec les liens sémantiques seulement puisqu'elle ne considère que les peuplements qui sont également liés sur le plan géométrique par une surface commune. De plus, il faut également tenir compte du fait que cette moyenne inclut le cas ou le volume obtenu est zéro parce que le peuplement ne respecte pas l'attribut descriptif spécifié pour la requête (par exemple, dans ce cas, le résultat d'une requête est nul si le peuplement ne comporte pas de sapins). Comme dans le cas des requêtes avec liens sémantiques seulement, les résultats du modèle MD, combinés à l'approche géométrique, sont plus faibles que ceux du modèle redéfini car ils négligent la similarité entre les propriétés. Cependant, l'écart moyen entre les deux séries de résultats est plus faible que dans le cas précédent compte tenu du fait que moins de peuplements sont impliqués dans le résultat, chaque peuplement étant associé à un écart supplémentaire.

L'importance de l'écart entre les résultats entre le modèle Matching Distance plus la géométrie et notre approche géosémantique indique l'importance de l'apport de la redéfinition du modèle de similarité sémantique puisque dans le cas des requêtes avec liens sémantiques, et les résultats montent également l'importance de tenir compte simultanément de l'évolution sémantique et géométrique pour obtenir des résultats acceptables pour les requêtes temporelles. Dans l'ensemble, les résultats obtenus à l'aide de notre approche montrent que la méthode proposée permet d'améliorer la réponse aux requêtes temporelles.

6. 6 Conclusion

Ce chapitre a permis de démontrer la force de notre approche. Cette dernière est composée de plusieurs phases : premièrement, elle comporte un algorithme de rétablissement de liens sémantiques au moyen d'un modèle de similarité sémantique redéfini à la fois générique et adapté aux données du domaine forestier. Puis, notre approche intègre un algorithme de rétablissement de liens géométriques au moyen du module Oracle et de l'indexation QuadTree. La fusion de ces deux approches sémantiques et géométriques produit une approche géosémantique intégrée pour le traitement des requêtes spatio-temporelles.

Les résultats des tests d'évaluation du modèle de similarité sémantique montrent que le modèle redéfini améliore à la fois le rappel, c'est-à-dire la capacité du modèle à repérer les instances similaires, ainsi que la précision, qui mesure la capacité du modèle à identifier les couples d'instances qui sont effectivement similaires.

Les résultats des requêtes temporelles montrent une amélioration nette des possibilités, car d'une part, elles permettent de fournir des résultats pour l'évolution des mesures liées aux peuplements, alors qu'a priori, aucune réponse n'était possible ou les résultats obtenus ne tenaient compte ni de l'évolution sémantique et ni de l'évolution géométrique. De plus, l'analyse comparative réalisée sur les résultats des requêtes montre que l'écart entre les résultats obtenus à partir du modèle de similarité original (modèle Matching Distance) et ceux qui furent obtenus à partir du modèle de similarité redéfini rend compte de l'apport de l'approche pour donner des réponses cohérentes aux requêtes spatio-temporelles.

Conclusion et recommandations

L'évolution de la structure multidimensionnelle des cubes de données géospatiales se produit lorsque sont effectuées des mises à jour du schéma de dimensions, du schéma des instances de la dimension ou des faits (ajout ou suppression de fait). L'évolution de la structure se produit également lorsqu'une même réalité, par exemple un territoire, est représentée à plusieurs époques par un cube de données géospatiales. Dans ce second cas, les liens sémantiques et géométriques entre les différents cubes de données géospatiales ne sont pas conservés et par conséquent les résultats des requêtes spatio-temporelles sont faussés ou impossibles à obtenir. L'évolution de la structure multidimensionnelle, parce qu'elle affecte les résultats des requêtes temporelles, entrave le processus de prise de décision mené par les utilisateurs. Actuellement, les solutions apportées à la problématique de l'évolution de structure ne considèrent que le premier cas où les relations entre les différents cubes sont connues puisque l'évolution est réalisée par l'administrateur lors des mises à jour. De plus, elles ne considèrent pas explicitement ou simultanément l'évolution sémantique et géométrique.

Ce mémoire s'intéresse à la problématique de l'évolution sémantique et géométrique de la structure multidimensionnelle. L'approche géosémantique développée dans le cadre de ce mémoire a pour objectif principal de gérer l'évolution de la structure en rétablissant, dans un premier temps, les liens sémantiques et géométriques entre les membres du schéma des instances des dimensions de différents cubes de données géospatiales, puis d'améliorer les résultats des requêtes spatio-temporelles. L'approche géosémantique a été élaborée en tenant compte des données forestières de la forêt de Montmorency, mais vise d'abord à constituer une méthode de gestion générale de l'évolution sémantique et géométrique dans les cubes de données géospatiales pouvant s'adapter à d'autres contextes.

Le développement de l'approche s'est déroulé en plusieurs étapes. Nous avons effectué une étude des solutions apportées à la problématique de l'évolution de la structure multidimensionnelle. Cette revue de littérature nous a permis d'identifier différents types d'évolution de la structure multidimensionnelle et de rendre compte de certaines faiblesses

des solutions proposées, en particulier qu'elles ne gèrent ni l'évolution sémantique et ni l'évolution géométrique et qu'elles ne considèrent pas que le rétablissement de liens entre les différents cubes de données géospatiales doit faire partie d'une solution complète pour la gestion de l'évolution de la structure multidimensionnelle. Nous avons alors envisagé comme objectif spécifique de rétablir les liens sémantiques et géométriques. Pour la problématique du rétablissement de liens sémantiques, nous avons exploré, dans le chapitre 2, le domaine des ontologies qui permet de représenter les connaissances d'un domaine par la définition des concepts et des relations. Nous avons établi un parallèle entre l'évolution de la structure multidimensionnelle et l'évolution des ontologies. Au niveau ontologique, plusieurs approches gèrent la problématique de l'hétérogénéité entre les ontologies et de leur évolution par des méthodes de mapping entre les ontologies. Parmi celle-ci, plusieurs emploient un modèle de similarité sémantique pour rétablir les relations entre les concepts de deux ontologies, c'est-à-dire pour effectuer le mapping entre les ontologies. En nous inspirant de ces approches, nous avons proposé un modèle de similarité sémantique pour rétablir les liens sémantiques entre les différents cubes de données géospatiales. Plusieurs modèles de similarité sémantique ont d'abord été étudiés. Le modèle proposé se base sur un modèle conçu pour établir l'interopérabilité entre les classes d'entités spatiales appartenant à différentes ontologies. Il a été adopté en raison de sa capacité en prendre en compte différents modes de représentation des classes d'entités spatiales : leurs propriétés, qui peuvent être de différentes natures, leur position dans le graphe de l'ontologie, ainsi que le contexte d'évaluation de la similarité sémantique qui s'adapte aux besoins de l'utilisateur. Malgré ses avantages par rapport aux autres modèles étudiés, nous avons estimé que ce modèle pouvait être amélioré afin d'être plus précis, c'est-à-dire en prenant compte du plus fin niveau de définition des concepts, et également afin d'être plus général, c'est-à-dire en prenant compte du fait que les propriétés des concepts peuvent être de différentes natures: des mots, mais également des textes ou des domaines de valeurs. Finalement, le modèle a également été amélioré en intégrant une mesure spécifiquement conçue pour évaluer la similarité sémantique entre les membres des niveaux supérieurs dans la structure hiérarchique. Ainsi, le modèle de similarité sémantique élaboré pour rencontrer notre premier objectif spécifique et présenté dans le chapitre 3 s'adapte à notre contexte, c'est-à-dire à la structure spatiale des dimensions des cubes de données géospatiales, mais

constitue également une amélioration générale d'un modèle de similarité sémantique pouvant être adopté dans d'autres domaines.

Dans le chapitre 4, nous nous sommes intéressés aux approches géométriques pour le rétablissement de liens géométriques entre les membres du schéma des instances de la dimension spatiale des différents cubes de données géospatiales, c'est-à-dire pour l'atteinte du second objectif spécifique. L'étude des données géométriques issues des inventaires de la forêt de Montmorency avec lesquelles nous avons testé notre approche nous a permis de constater qu'en général, l'évolution géométrique était considérable, c'est-à-dire que la forme des polygones forestiers a été fortement modifiée d'un inventaire à l'autre. Le rétablissement de liens géométriques devait donc se fonder sur une représentation géométrique commune à toutes les cartes forestières de chaque époque plutôt que sur un appariement géométrique entre les polygones des différentes cartes. Pour ce faire, nous avons étudié quelques méthodes d'indexation de données spatiales. La création de la représentation commune a été effectuée au moyen de la méthode d'indexation *QuadTree* du module *Oracle Spatial*. L'arbre d'indexation *QuadTree* permet de représenter les polygones par un ensemble de cellules invariables entre les différentes époques et s'avère également approprié pour représenter des polygones de forme variable, puisque la taille de cellules permet de s'adapter à la distribution de l'information spatiale. Le rétablissement de liens géométriques a été effectué au moyen du taux d'inclusion liant les polygones de différentes époques et sur la base des tables d'indexation produites par la méthode *QuadTree*.

Le troisième objectif spécifique a été réalisé par le développement de l'approche géosémantique intégrée qui combine les résultats des rétablissements de liens sémantiques et géométriques, représentés sous forme matricielle, avec une méthode de transformation matricielle permettant de traiter les requêtes spatio-temporelles dans les cubes de données géospatiales. La méthode de transformation matricielle représente, d'une part, les cubes de données sous forme matricielle et, dans un deuxième temps, permet de transformer les mesures issues d'un premier cube de données dans la structure multidimensionnelle associée à un second cube par les matrices qui contiennent les liens sémantiques et géométriques entre les membres du schéma des instances de la dimension entre des cubes

différents. L'approche géosémantique permet ainsi d'atteindre l'objectif principal qui est d'améliorer les résultats des requêtes temporelles dans les cubes de données géospatiales affectés par l'évolution sémantique et géométrique de leur structure. Nous soulignons également que cette méthode présente des résultats encourageants, puisque, a priori, dans la plupart des cas, il était même impossible d'obtenir des résultats aux requêtes temporelles. En effet, les membres des dimensions spatiales du cube, qui sont des zones forestières, n'étaient pas liés entre eux d'une époque à l'autre.

Dans le dernier chapitre, nous avons testé et évalué l'applicabilité de notre approche dans le contexte forestier. L'implantation constitue un apport supplémentaire par rapport aux objectifs fixés. L'application développée montre que l'approche géosémantique est applicable concrètement. Premièrement, les tests de performance réalisés pour le modèle de similarité sémantique montrent que le modèle proposé permet effectivement d'améliorer la précision par rapport au modèle original (le modèle Matching Distance), et permet également d'améliorer le rappel, qui mesure la capacité du modèle à repérer les instances qui sont similaires. Le prototype test conçu en intégrant SOLAP, l'extension spatiale de l'outil OLAP, aux résultats du rétablissement des liens sémantiques obtenus avec l'approche sémantique montre également que les liens sémantiques peuvent aider l'utilisateur à analyser de manière rapide et visuelle l'évolution des propriétés sémantiques des peuplements selon différents points de vues, soit en observant la similarité entre les essences des peuplements forestiers, leurs attributs ou leur fonction, leur voisinage, etc., ce qui constitue un apport supplémentaire par rapport aux objectifs fixés. Finalement, les résultats des réponses aux requêtes montrent que l'approche permet de donner des résultats cohérents et qu'il est nécessaire de tenir compte à la fois de l'évolution sémantique et géométrique.

Nous pouvons donc affirmer que les objectifs spécifiques fixés pour cette recherche, qui sont de rétablir les liens sémantiques et géométriques et de développer une approche géosémantique, ont été atteint, et que notre objectif principal a également été atteint puisque la combinaison des approches développées permet de répondre et d'améliorer la réponse aux requêtes spatio-temporelles.

Plusieurs points peuvent être abordés pour améliorer l'approche développée. Premièrement, l'approche fut testée avec un nombre restreint de données, il aurait donc été intéressant de l'évaluer sur un volume de données plus important pour ainsi tester un plus grand nombre de requêtes. Sur le plan théorique, notre approche demeure flexible car le modèle de similarité sémantique permet d'intégrer différents types de propriétés pour lesquelles il est possible de définir une mesure de similarité sémantique. Cependant, il aurait été intéressant de pouvoir tester l'approche géosémantique dans un autre domaine d'application où les concepts possèdent différentes propriétés, par exemple dans le domaine de la santé, afin d'en montrer concrètement la flexibilité. De plus, le cadre théorique de notre approche aurait pu être amélioré par le développement d'un métamodèle temporel. Puis, il serait intéressant, pour améliorer les réponses aux requêtes, d'intégrer d'autres méthodes qui puissent automatiser la détermination des poids non seulement par rapport au contexte mais également au profil de l'utilisateur, et puis d'intégrer des algorithmes d'optimisation qui puissent améliorer le calcul des poids qui pourrait se baser sur des méthodes différentes dépendamment du contexte. Ainsi, nous avons évalué qu'en général, le principe de la ressemblance donne des meilleurs résultats par rapport au principe de variabilité, puisqu'il est plus adapté au contexte forestier, mais nous pourrions développer un algorithme relevant du domaine de la fouille de données (Data mining) qui détecterait, dans un contexte arbitraire, lequel des principes (entre la ressemblance et la variabilité) s'avère plus pertinent pour la détermination des poids.

De plus, en ce qui concerne l'approche géométrique, il serait intéressant d'explorer le mode de représentation matricielle des données, telles que la méthode de *Voronoi* et la représentation hexagonale afin d'optimiser les résultats du traitement des données spatiales. Nous avons également constaté que dans le cas où l'évolution géométrique entre les polygones est faible, comme c'est le cas entre les polygones de l'inventaire de 1984 et de celui de 1992, il serait intéressant d'utiliser une méthode vectorielle pour obtenir plus de précision. Il serait également intéressant, dans l'avenir, de comparer les résultats obtenus avec notre approche avec ceux qui ont été obtenus avec l'approche de Miquel, M., Y. Bédard & A. Brisebois, laquelle vise également à prendre en compte l'évolution entre les inventaires forestiers. Nous recommandons finalement de poursuivre le couplage de notre approche aux outils SOLAP pour pouvoir effectuer des analyses géostatistiques.

Bibliographie

Ahmed T.O., Miquel M., Laurini R. (2004) *Continuous Data Warehouse: Concepts, Challenges and Potentials*. Proc. 12th Int. Conf. on Geoinformatics − Geospatial Information Research: Bridging the Pacific and Atlantic University of Gävle, Sweden, p. 7-9.

Allen, J. F. (1983) *Maintaining Knowledge about Temporal Intervals.* (Tome 26 (11).Communication of the ACM, p. 832 – 843.

Amini, M-R. (2001) *Apprentissage automatique et recherche d'information : application à l'extraction d'information de surface et au résumé de texte*. Thèse de doctorat, Paris, France, Université Paris 6. 216 pages.

Bakillah, M., Mostafavi, M.A., Bédard, Y. (2006) A Semantic Similarity Model for Mapping between Evolving Geospatial Data Cubes. OTM Workshop 2006, LNCS 4278, p. 1658-1669.

Baziz, M. (2005) *Indexation conceptuelle guide par ontologie pour la recherche d'information*. Thèse de doctorat, Université Paul Sabatier, 234 pages.

Bębel B., Wrembel R., Królikowski Z. (2004) *Transaction Concepts of Data Warehouses*. Pro Dialog (Polish Information Processing Society Journal), no. 18, ISBN 83-89529-00-9, ISSN 0867-6011, p. 143-149.

Bédard Y., Merrett T., Han J. (2001) *Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery*. Chapitre du livre *Geographic Data Mining and Knowledge Discovery* édité par H. Miller and J. Han, Research Monographs in GIS series, édité par Peter Fisher and Jonathan Raper, Taylor & Francis.

Bédard Y., Proulx M.-J., Rivest S. (2005) *Enrichissement du OLAP pour l'analyse géographique: exemples de réalisations et différentes possibilités technologiques*. Première journée francophone sur les entrepôts de données et l'analyse en ligne, Lyon.

Bench-Capon, T, Malcolm, G. (1999) *Formalising Ontologies and their Relations*. Proceedings of the 16th International Conference on Database and Expert Systems Applications, p. 250-259.

Blaschka, M. (1999) FIESTA: A Framework for Schema Evolution in Multidimensional Information Systems. In Proceedings of the 6th CAISE Doctoral Consortium, Heidelberg, Germany.

Barwise, J., Seligman, J. (1997) *Information Flow: the Logic of Distributed Systems*. Cambridge University Press.

Body M., Miquel M., Bédard Y., Tchounikine A. (2002) *A Multidimensional and Multiversion Structure for OLAP Applications*. ACM Fifth International Workshop on Data Warehousing and OLAP, Proceedings, p. 1-6.

Body M., Miquel M., Bédard Y., Tchounikine A. (2003) *Handling Evolutions in Multidimensional Structures*. IEEE 19th Int. Conf. on Data Engineering (ICDE), March 5-8, Bangalore, India, Bangore, Inde.

Bounif H., Spaccapietra S. (2003) *Predictive Database Schema Evolution*. Workshop on Multimodal Interaction and Related Machine Learning algorithms, p. 647-651

Bittner S. (2001) An Agent-Based Model of Reality in a Cadastres. Thèse de doctorat. Technischen Universität Wien, 219 pages.

Brickley, D., Hunter, J., Lagoze, C. (1999) ABC: *A Logical Model for Metadata Interoperabilit,* p.116- 147

Brodeur, J. (2004) *Interopérabilité des Données Géospatiales : Élaboration du Concept de Proximité Géosémantique.* Thèse de doctorat, Université Laval, 267 pages.

Cabibbo L., Torlone R. (1997) *Querying Multidimensional Databases*. In Proceeding of the 6th international workshop on Database Programming Languages, East Park, Colorado, USA, p. 253-269.

Carbo, R., Domingo, L. (1987). *LCAO-MO similarity measures and taxonomy*. Int. J. Quant. Chem. 32, p. 517-545.

Carbo-Dorca, R., Besalu E. (1998) *A general survey of molecular quantum similarity*.J. Mol. Struct. - Theochem 451(1-2), p. 11-23.

Cabibbo, L., Torlone R. (1997) *Querying Multidimensional Databases*. In Proceeding of the 6th international workshop on Database Programming Languages, East Park, Colorado, USA, p. 253-269.

Carron, P-Y, (1998), *Étude du potentiel OLAP pour supporter l'analyse spatio-temporelle. Mémoire de maîtrise, Université Laval.*

Champin, P.-A., Solnon, C. (2003) *Measuring the Similarity of Labeled Graphs*. LNCS, Vol. 2689, p. 80-95.

Chandrasekaran, B., Josephson, J.R., Benjamin, V.R.(1999) *What are Ontologies, and why do we need them ?* Intelligent Systems and their Application, Volume 14, Issue 1, p.20-26.

Charlet, J., Bachimont, B., Troncy, R. (2003). *Ontologies pour le Web sémantique*. In J. Charlet, P. Laublet et C. Reynaud (Eds.), *Web sémantique*: Action_spécifique_32_CNRS/STIC.

Chaudhuri S., Dayal U. (1997) *An Overview of Data Warehousing and OLAP Technology*, ACM SIGMOD Record, 26(1).

Chalupsky, H. (2000) *Onto-Morph: A translation system for Symbolic Logic.* In Anthony G Cohn, Fausto Giunchiglia and Bart Selman, editors, KR2000: Principles of Knowledge Representation and Reasoning, San Francisco, CA, p. 471-482.

Compatangelo, E, Meisel, H. (2002) *Intelligent support to knowledge sharing trough the articulation of class schema.* Proceeding of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems.

Doan, A., Madhavan, J., Domingos, P., Halevy, A. (2002) *Learning to map between ontologies on the Semantic Web*. In the 11th International WWW Conference, Hawaii, US.

Eder, J., Koncilia, C. (2001) *Changes of Dimension Data in Temporal Data Warehouses*. Y. Kambayashi, W. Winiwarter, M. Arikawa. (Eds): Da WaK 2001, LNCS 2114, p. 284-293.

Eder, J., Koncilia, C. (2004) *Modelling Changes in Ontologies*. OTM Workshops 2004, LNCS, 3292, p. 662-673.

Egenhofer, M., Franzosa, R.D. (1991) *Point-Set Topological Spatial Relations*. International Journal of Geographical Information Systems, 5(2), p.161–174.

E. Morin (1999) *Extraction de lien sémantique entre termes à partir de corpus de textestechniques*. Thèse de doctorat, Institut de recherche en informatique de Nantes, 215 pages.

Erwig M., Güting R.H., Schneider M. Vazirgiannis M.(1999) *Spatio-Temporal Data Types : An Approach to Modelling and Querying Moving Objects in Databases*. Geoinformatica, Vol 3.No 3:p. 269-296, 1999.

Fagan, J.L. (1989) *The effectiveness of a nonsyntactic approach to automatic phrase indexing or document retrieval*. Journal of the American Society for Information Science, 40(2), p. 115–132.

Fernandez, M., Gomez-Perez, A., Juristo, N. (1997) *Methontology: From Ontological Art Toward Ontological Engineering.* Paper presented at the Spring Symposium Series on Ontological Engineering, AAAI97, Stanford, USA.

Gajda, E.M. (2003) *Concepts and methodological framework for spatio-temporal data warehouse design*. Thèse de doctorat.Université Libre de Bruxelles. p.107

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. *(2003) Sweetening WordNet with DOLCE.* AI Magazine, 24(3), p. 13-24.

Golfarelli, M., LechtenBürger, J., Rizzi, S., Vossen, G. (2004) *Schema Versioning in Data Warehouse*. S. Wang et al. (Eds.): ER Workshops 2004, LNCS 3289, p. 415–428.

Gomez-Perez, A. (1999) *Ontological Engineering: A state of the art*. Madrid: Facultad de Informatica, Universidad Politecnica de Madrid.

Grandi F., Mandreoli F. (2002) *A Formal Model for Temporal Schema Versioning in Object-Oriented Databases*. A Time Center Technical Report.

Gruber, T.R.(1995) *Toward Principles for the Design of Ontologies used for Knowledge Sharing*. International Journal of Human–Computer Studies, 43, p. 907–928.

Guarino, N. (1998) *Formal Ontology and Information Systems*. In Proceedings of Formal Ontology in Information Systems. Amsterdam, IOS Press, p.3-15.

Guerrero, E.I.B. (2002) *Infrastructure Adaptable pour l'évolution des entrepôts de données*. Thèse de doctorat, Université Joseph Fourier.

Hurtado, C.A., Mendelzon, A.O., Vaisman, A.A. (1999) *Maintaining Data Cubes under Dimension Updates*. In 15th Int. Conf. on Data Engineering, p 346–355.

Hjaltason, G.R., Samet, H. (2002) *Speeding up Construction of PMR Quadtree-Based Spatial Indexes*. The International Journal on Very Large Data Bases, Vol. 11(2), p. 109 – 137.

Hurtado, C.A., Mendelzon, A.O., Vaisman, A.A. (1999) *Maintaining Data Cubes under Dimension Updates*. In 15th Int. Conf. on Data Engineering, p. 346–355.

Inmon, W.H. (1996) *Building the Data Warehouse*. John Wiley & Sons.

Jensen C.S., Clifford J., Gadia S.K., Hayes P., Jajodia S. et al. (1998) *The Consensus Glossary of Temporal Database Concepts - February 1998 Version*. In O. Etzion, S. Jajodia, and S. Sripada,editors, *Temporal Databases - Research and Practice*, Springer-Verlag. LNCS No. 1399, p. 367–405.

Jiang, J., Conrath, D. (1997) *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. In: International Conference on Computational Linguistics. (ROCLING X), Taiwan, p. 19-35.

Jiang, J.J., Conrath, D.W. (1998) *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. In Proceedings of the International Conference on Research in Computational Linguistic, Taiwan.

Kalgoflou, Y., Shorlemmer, M (2002) *Information flow-based Ontology Mapping* In: On the Move to Meaningful Internet Systems 2002; CoopIS, DOA, and ODBase lectures notes in Computer Science 2519, Springer, p.1132-1151.

Kalfoglou, Y, Schorlemmer, M. (2003) *Ontology Mapping: the State of the Art*. The Knowledge Engineering Review, Vol. 18 (1), p. 1-31.

Kimball, R. (1996) *The Data Warehouse Toolkit*. J. Wiley and Sons Inc.

Kimball, R. (2002a) *OLAP and ROLAP are a continuum,* not competitors by Joy Mundy.

Klein, M. (2001) *Combining and relating ontologies: an analysis of problems and solution*. In IJCAI-2001 Workshop on Ontologies and Information Sharing, p53-62, Seattle, WA.

Klein, M., Kiryakov, A., Ognyanov, D., Fensel, D. (2002) *Ontology Versioning and Change Detection on the Web.* In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), 2002.

Klein, M., Noy, N. (2003). *A component-based framework for the ontology evolution.* Paper presented at the Workshop on Ontologies and Distributed Systems, IJCAI-2003, Acapulco, Mexico.

Knappe, R., Bulskov, H., Andreasen, T. (2003) *On Similarity Measures for Content-Based Querying.* In O. Kaynak, editor, Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA'03), p. 400–403, Instanbul, Turkey.

Koncilia, C. (2003) A *bi-temporal data warehouse model*. In Proc. of Short Papers of the 15th Int. Conf. on Advanced Information Systems Engineering, p. 77-80.

Lacher, M., Groh, G. (2001) *Facilitating the Exchange of Explicit Knowledge through ontology mapping*. Proceedings of the14th International FLAIRS Conference.

Lefebvre, P. (2000) *La Recherche d'Information – du Texte Intégral au Thesaurus*. Hermes Science, Paris.

Li, Y., Bandar, Z.A., McLean, D. (2003) *An Approach for Measuring Semantic Similarity between Words using Multiples Information Sources.* IEEE Transactions on Knowledge and Data Engineering, 15(4), p. 871, 882.

Lin, D. (1993) *Principle-Based Parsing Without Overgeneration*. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93), p.112–120, Columbus, Ohio.

Mendelzon, A.O., Vaisman, A.A. (2000) *Temporal Queries in OLAP*. Proceedings of 26th International Conference on Very Large Data Bases - VLDB 2000, Cairo (Egypt).

Morzy T., Wrembel, R. (2004) *On querying Versions of Multiversion Data Warehouse*. DOLAP'04, Washington, DC, USA.

Mostafavi, M. A., (2006) Semantic similarity assessment in support of geospatial data integration. The Seventh International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Portugal, pp. 685-693.

McGuinness, DL, Fikes, R, Rice, J, Wilder, S. (2000) *An Environment for Merging and Testing Larges Ontologies*. In: Cohn, AG, Giunchiglia, F, Selman, B. (eds), KR2000: Principles of Knowledge Representation and Reasoning, San Francisco, p. 483-493.

Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R. (2003) *Ontologies for Enterprise Knowledge Management*. Intelligent Information Processing.

Malinowski E., Zimányi, E. (2004) *Representing Spatiality in a Conceptual Multidimensional Model*. In Proc.GIS.

Malinowski E., Zimànyi, E. (2005) *Hierarchies in a multidimensional model: from conceptual modeling to logical representation*. Accepted for publication in Data & Knowledge Engineering.

Malinowski E., Zimányi, E. (2004) *Representing Spatiality in a Conceptual Multidimensional Model*. In Proc.GIS 2004.

Malinowski E., Zimànyi, E. (2005) *Hierarchies in a multidimensional model: from conceptual modeling to logical representation*. Accepted for publication in Data & Knowledge Engineering.

Malinowski E., Zimànyi, E. (2006) *A Conceptual Solution for Representing Time in Data Warehouse Dimensions*. Third Asia-Pacific Conference on Conceptual Modelling (APCCM2006), Hobart, Australia.

Manning, C.D., Schütze, H. (1999) *Foundation of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

McBrien P., Poulovassilis A. (2002). *Schema Evolution in Heterogeneous Database Architectures, A Schema Transformation Approach*. Proc. CAiSE'02, Toronto, May 2002, LNCS 2348, p. 484-499.

Mendelzon A.O., Vaisman A.A. (2000) *Temporal Queries in OLAP*. Proceedings of 26th International Conference on Very Large Data Bases - VLDB 2000, Cairo (Egypt).

Miquel, M., Y. Bédard & A. Brisebois (2002) *Conception d'entrepôts de données géospatiales à partir de sources hétérogènes, exemple d'application en foresterie*, Ingénierie des Systèmes d'information,Vol. 7, No. 3, p. 89-111

Morzy T., Wrembel R. (2004) *On querying Versions of Multiversion Data Warehouse*. DOLAP'04, Washington, DC, USA.

Najib, F., Godin, R., Missaoui, R., David, S., Plante, P. (1996) *Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte*. Canadian Journal of Information and Library Science / Revue l'information et de bibliothéconomie, 1996, 21(1), p. 1-21.

Nascimento, M., Silva, A., J.R.O., Theodoridis, Y. (1999) *Evaluation for Access Structures for Discretely Moving Points*. In Proc. of the Int. Workshop on Spatio-Temporal Database Management, Edinburgh, UK, p. 171-188. Springer-Verlag.

Nascimento, M., Silva, A., J.R.O. (1998) *Toward Historical R-Tree*. In Proc. of the ACM Symp on Applied Computing, Atlanta, GA, pp. 235-240. ACM Press.

Niles, I., Pease, A. (2001) *Toward a Standard Upper Ontology*. In the 2nd International Conference on Formal Ontology in Information Systems, Ogunquit, Maine.

Noy, N., Klein, M.(2003) *A component-based framework for ontology evolution.* In Proceedings of the Workshop on Ontologies and Distributed Systems (IJCAI'03).

Noy, N.F., Musen, M.A. (2003) The *PROMPT Suite: Interactive Tools for Ontology Merging and Mapping*. International Journal of Human-Computer Studies, 59(6): 983-1024.

Orzanco, M. G. *Fusion de l'information forestière issue de différentes sources pour améliorer la fiabilité locale d'une carte forestière*. Thèse de doctorat, Université Laval, 2005.

Ouksel, A.M., Sheth, A.(1999) *Semantic Interoperability in Global Information Systems : Abrief Introduction to the Research Area and the Special Section*. Sigmod Record, 28(1), p. 5-12.

Pedersen, T.B., Jensen, C.S (1999) *Multidimensional Data Modeling for Complex Data.* In Proceedings of the International Conference on Data Engineering -ICDE'99.

Pederson, T.B., Jensen, C.S., Dyreson C.E. (2001) A Foundation for Capturing and Querying complex multidimensional Data. Information Systems 26 (2001) p. 383–423.

Pestana G., Da Mira, M.M. (2005) *Multidimensional Modeling Based on Spatial, Temporal and Spatio-Temporal Stereotypes*. ESRI International User Conference July-25-29, San Diego Convention Center, San Diego, Californie.

Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989) *Developement and Application of a Metric on Semantic Nets*. IEEE Transactions on Systems, Man and Cybernetics 19(1), p. 17-30.

Rebout, C. (1998) *Adaptation d'une base de données pour une application SOLAP pour l'aide à l'aménagement intégré des ressources forestières*. Rapport de DESS, Université Joseph Fourier, Grenoble, France.

Rigaux, P., Scholl, M., Voisard, A. (2000) *Spatial Databases*. Morgan Kaufmann.

Rivest S., Bédard Y., Marchand P. (2001) *Toward better Support for Spatial Decision Making : Defining the Characteristics of Spatial On-Line Analytical Processing*. Geomatica, Vol 55 (4), p. 539-555.

Rodriguez, A. (2000) *Assessing Semantic Similarity among Spatial Entity Classes*. Thèse de doctorat, University of Maine, p.168

Salton, G., McGill, M. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Seco, N., Veale, T.,Hayes, J. (2004) *An Intrinsic Information Content Metric for Semantic Similarity in WordNet*. In Proceedings of ECAI 2004, the 16th European Conference on Artificial Intelligence.

Shimm J.P., Warkentin M., Courtney J., Power D., Sharda R., Carlsson Ch. (2002) *Past, Present and Future of Decision Support Technology*. In Decision Support Systems, Elsevier Science B.V., Vol. 33(2), P. 111 - 126

Simoni, J.L.(2000) *Accès à l'information à l'aide d'un graphe de termes construit automatiquement (Intégration de l'interrogation et de la navigation)*. Thèse de doctorat, Université Paris 7.

Snodgrass R. (1995) *The TSQL2 Temporal Query Language*. Kluwer Academic Publishers.

Sure, Y., Staab, S., et Studer, R. (2004) *On-To-Knowledge Methodology (OTKM)*. In S. Staab et R. Studer (Eds.), *Handbook on Ontologies* Springer Verlag, p. 117-132.

Terenziani P., Snodgrass R.T., Bottrighi A., Torchio M., Molino G. (2006) *Extending Temporal Databases to deal with Telic/Atelic Medical Data*. Time Center Technical Report.

Theodoridis, Y.,Vazirgiannis, M., Sellis, T. (1996) *Spatio-temporal Indexing for Large Multimedia Applications*. In Proc. of the 3rd IEEE Conf. on Multimedia Computing and Systems, p. 441-448.

Tversky, A. (1977) *Features of Similarity*. Psychological Review 84(4): 327-352.

Vaisman, A.A. (2001) Updates, View Maintenance and Time Management in Multidimensional Databases. Thèse de Doctorat. Université de Buenos Aires.

Visser, P.R.S., Jones, D.M., Bench-Capon, T.J.M, Shave, M.J.R. (1997) *An Analysis of Ontological Mismatches: Heterogeneity versus Interoperability*. In: AAAI 1997 Spring Symposium on Ontological Engineering, Stanford, USA, p 164- 172.

Xu, X., Han, J., Lu, W. (1990) *RTree : An Improved R-Tree Index Structure for Spatio-Temporal Database*. In Proc. of the 4th Int. Symp. On Spatial Data Handling, Zurich, Switzerland, p. 1040-1049.

Zhang D., Tsotras V. (2001) *Improving Min/Max aggregation over Spatial Object*. In Proc. GIS, 2001.

Zhang, J., Goodchild, M. (2002) *Uncertainty in Geographical Information*. Taylor & Francis, London. 266 pages.

Zurita, V. (2004) *Semantic-based Approach to Spatial Data Sources Integration*. Thèse de doctorat, Universitat Politècnica de Catalunya.

Annexe A

Tableau 1A: Peuplements 73 (les codes référent à des définitions et des domaines de valeurs des peuplements donnés dans les tableaux 4A à 7A).

Peuplement 73	essence	âge	perturbation	densité	hauteur	APPELAT73	fonction
1	S	2		2	4	S B4 j	reserve ecologique
2	S	2	cp	3	3	S C3 j cp	bloc experimental
3	S	2		2	3	S B3 j	production
4	S	2		2	3	S B3 j	production
5			friche			friche	production
6	S	4		2	2	S B2 mr	production
7	S	4		2	2	S B2 mr	production
8	S	2	cp	3	3	S C3 j cp	production
9	S	2	cp	4	3	S D3 j cp	production
10			ct			ct	reserve ecologique
11	S	2	cp	4	3	S D3 j cp	bloc experimental
12	S	2		2	3	S B3 j	bloc experimental
13			aulnaie			aulnai	production
14	S	2		2	4	S B4 j	production
15	S	4	cp	3	2	S C2 mr cp	reserve ecologique
16			friche			friche	bloc experimental

Tableau 2A: Peuplements 84 (les codes référent à des définitions et des domaines de valeurs des peuplements donnés dans les tableaux 4A à 7A).

Peuplement 84	essence	DH84	âge	densité	hauteur	perturbation	fonction
1	S(S)	B4	300	2	4		terrain non productif
2	S(S)	C3	700	3	3	cp	production
3	S(S)	B3	700	2	3		production
4			0			ct	bloc expérimentale
5	S(S)	B3	700	2	3		bloc expérimental
6	S(S)	B3	700	2	3		production
7	S(S)	D3	700	4	3	cp	bloc expérimental
8	S(S)	B3	700	2	3		bloc expérimental
9	S(S)	B3	700	2	3		production
10	S(S)	A4	300	1	4		terrain non productif
11	R	6	100		6	ct	terrain non productif
12	S(S)	B3	700	2	3		terrain non productif
13	S(S)	B3	700	2	3		terrain non productif
14	S(S)	A4	300	1	4		terrain non productif
15	S(S)	A3	500	1	3		terrain non productif
16	S(S)	A3	500	1	3		production
17	S(S)	B2	900	2	2		production
18	S(S)	B2	900	2	2		production
19	S(S)	C3	700	3	3	cp	terrain non productif
20	S(S)	A3	500	1	3		production
21	S(S)	B2	900	2	2		production
22	S(S)	A4	500	1	4		bloc expérimental
23	S(S)	A4	500	1	4		bloc expérimental
24	S(S)	B3	700	2	3		bloc expérimental
25	BbS(F)	B4	300	2	4		bloc expérimental
26	S(S)	B3	900	2	3		bloc expérimental
27	S(S)	A4	300	1	4		bloc expérimental
28	S(S)	B3	900	2	3		production
29	S(S)	B3	700	2	3		production
30	BbS(R)	A4	500	1	4		bloc expérimental
31	S(S)	C3	900	3	3	cp	bloc expérimental
32	BbS(R)	B2	900	2	2		production
33	BbS(R)	A4	500	1	4		bloc expérimental
34	R	6	100		6	ct	bloc expérimental
35	S(S)	A4	300	1	4		bloc expérimental
36	S(S)	B3	700	2	3		production

37			0			ct	production
38	BbS(R)	A4	500	1	4		production
39	S(S)	C3	700	3	3	cp	bloc expérimental
40	S(S)	C3	700	3	3	cp	bloc expérimental
41			0			ct	bloc expérimental
42	S(S)	A4	300	1	4		bloc expérimental
43	S(S)	A4	300	1	4		production
44	S(S)	B3	700	2	3		bloc expérimental
45	S(S)	B3	700	2	3		terrain non productif
46	S(S)	D3	700	4	3	cp	terrain non productif
47	S(S)	C3	700	3	3	cp	terrain non productif
48	S(S)	B3	500	2	3		
49	S(S)	B4	300	2	4		
50	BbS(R)	B3	500	2	3		
51	S(S)	B3	700	2	3		
52	S(S)	B3	700	2	3		
53	S(S)	B4	300	2	4		
54	BbS(R)	A4	300	1	4		
55	S(S)	B3	500	2	3		terrain non productif
56	S(S)	B3	500	2	3		
57	S(S)	C3	700	3	3	cp	
58	BbS(R)	B3	500	2	3		
59	S(S)	B4	300	2	4		terrain non productif
60	S(S)	A4	300	1	4		
61	BbS(R)	A4	300	1	4		
62	S(S)	B3	500	2	3		
63			0			ct	
64	BbS(R)	B3	500	2	3		
65	S(S)	A4	300	1	4		

Tableau 3A: Peuplements 92 (les codes référent à des définitions et des domaines de valeurs des peuplements donnés dans les tableaux 4A à 7A).

Peuplement 92	HA92	pente	âge	densité	hauteur	essence	perturbation	fonction
2	4,6448	A	200					bloc expérimental
3	2,5434	A	30	1	4	SS		bloc expérimental
4	0,5568	A	200					bloc expérimental
5	8,896	B	1	4	6		ct	réserve écologique
6	0,9327	A	45	3	4	SE		réserve écologique
7	0,1176	D	30	2	4	SBb		réserve écologique
8	0,24	D	30	2	4	SBb		production
9	1,6033	A	30	2	4	SS		production
10	0,5745	D	30	2	4	SBb		production
11	3,8575	D	1	3	6		ct	réserve écologique
12	0,8081	A	45	3	4	EE		production
13	0,2608	A	30	1	4	SS		production
14	7,7003	C	1	3	6		ct	production
15	2,5161	B	15	2	5	SS	ct	production
16	0,9206	A	200					production
17	1,0873	A	30	2	4	SS		réserve écologique
18	0,9794	B	15	1	4	SBb		bloc expérimental
19	1,407	B	1	3	6		ct	bloc expérimental
20	0,4369	A	45	3	4	EE		bloc expérimental
21	1,0161	B	1	4	6		ct	réserve écologique
22	0,4551	B	30	1	4	SS		réserve écologique
23	0,4618	A	1	4	6		ct	réserve écologique
24	0,4955	E	30	1	4	SS		réserve écologique
25	1,8559	D	1				ct	réserve écologique
26	3,059	A	30	2	4	SS		réserve écologique

27	0,336	E	30	2	4	SBb		
28	1,589	B	30	1	4	SS		réserve écologique
29	2,78	E	90	4	3	BbS		
30	0,594	C	30	3	4	SS		
31	1,312	D	45	3	3	SBb		
32	0,196	D	30	2	4	SBb		réserve écologique
33	1,34	A	75	2	3	SS		
34	0,968	B	15	4	5	SS	ct	
35	0,557	B	90	4	3	BbS		
36	2,563	B	30	2	4	SS		
37	1,102	C	30	3	4	SS		
38	0,301	C	1				es	
39	0,851	B	75	4	3	SBb		
40	1,121	B	30	2	4	SS		
41	0,594	B	60	3	4	SE		réserve écologique
42	1,016	A	45	3	3	SS		bloc expérimental
43	1,345	B	75	4	3	SBb		bloc expérimental
44	0,727	A	500					
45	0,506	A	30	1	4	SS		
46	0,697	C	45	3	3	SS		
47	0,195	D	45	3	4	SE		
48	0,809	C	1	4	6		ct	
49	0,692	B	30	4	4	SS		
50	1,292	D	45	2	3	SBb		
51	0,792	B	30	2	4	SBb		
52	7,030	E	90	3	3	SS		
53	0,681	A	30	2	4	SBb		
54	1,857	A	1	4	6		ct	
55	1,482	D	75	2	3	BbS		
56	0,277	A	1	3	6		ct	
57	1,172	C	30	2	4	SS		
58	0,249	B	1	3	6		ct	
59	0,827	B	1	3	6		ct	réserve écologique
60	0,666	A	1	3	6		ct	
61	7,106	A	1	4	6		ct	
62	1,48	B	90	3	3	SS		
63	3,292	A	1	2	5	SS	es	
64	0,968	D	30	3	3	SBb		
65	0,501	C	1	3	6		ct	
66	1,074	B	60	3	4	SS		
67	1,231	A	30	2	4	SS		
68	3,056	B	1	3	6		ct	
69	1,598	D	75	2	3	SBb		
70	0,515	B	45	2	3	SS		
71	0,217	A	30	3	4	SS		
72	0,460	B	1	4	6		ct	réserve écologique
73	2,062	A	45	2	4	SS		
74	0,90	D	45	3	3	SBb		

Tableau 4A: Codes d'essences (tiré de Rebout, 1998)

Code essence	Inven-taire	Descriptif	Type d'essence	Classe	Essence
Fi	73	Feuillus d'essences intolérantes : jeune peuplement mélangé, dont le bouleau à papier et/ou les peupliers, seuls ou accompagnés d'une proportion variable d'érable rouge, occupent plus de 50% de la surface terrière de la partie feuillue ; est également classifié comme tel un peuplement mûr où la partie feuillue est composée des mêmes essences dans une proportion à peu près égale.	Feuillus	Feuillus classables	Feuillus intolérants
FiR(F)	73	Feuillus d'essences intolérantes avec résineux : un peuplement mûr où les feuillus occupent plus de 50% de la surface terrière.	Mélangés	Feuillus intolérants et résineux	Feuillus intolérants et résineux
Bb	73	Bétulaie à bouleaux blancs : un peuplement mur où le bouleau blanc occupe plus de 50% de la surface terrière de la partie feuillue.	Feuillus	Feuillus classables	Bouleaux blancs
E	73	Peuplement où l'épinette noire et/ou rouge occupe au moins 50 % de la surface terrière de la partie résineuse du peuplement	Résineux	Un résineux dominant	Épinettes
BbR(F)	73	Bétulaie à bouleaux blancs avec résineux : un peuplement mur où le bouleau blanc occupe plus de 50% de la surface terrière.	Mélangés	Bouleaux et résineux	Bouleaux blancs et résineux
BbR(R)	73	Sapinière à bouleaux blancs : un peuplement mûr où les résineux occupent plus de 50% de la surface terrière.	Mélangés	Résineux et bouleaux	Résineux et bouleaux blancs
FiR(R)	73	Feuillus d'essences intolérantes avec résineux :un peuplement mûr où les résineux occupent plus de 50% de la surface terrière.	Mélangés	Résineux et feuillus intolérants	Résineux et feuillus intolérants
S	73	Peuplement où le sapin et/ou l'épinette blanche occupent au moins 50 % de la surface terrière de la partie résineuse du peuplement	Résineux	Un résineux dominant	Sapins
BbS(R)	84	Sapinière à bouleaux blancs : un peuplement mûr où les résineux occupent plus de 50% de la surface terrière.	Mélangés	Résineux et bouleaux	Résineux et bouleaux blancs
BbS(F)	84	Bétulaie à bouleaux blancs avec résineux : un peuplement mur où le bouleau blanc occupe plus de 50% de la surface terrière.	Mélangés	Bouleaux et résineux	Bouleaux blancs et sapins
E(E)	84	L'épinette noire et/ou rouge occupe au moins 75% de la surface terrière de la partie résineuse du peuplement.	Résineux	Un résineux dominant	Épinettes
E(S)	84	L'épinette noire et/ou rouge occupe au moins 50% de la surface terrière de la partie résineuse du peuplement tandis que le sapin et/ou l'épinette blanche occupent au moins 25 % de cette même partie.	Résineux	Mélange de résineux	Épinettes et sapins
PeS(R)	84	Sapinière avec peuplier	Mélangés	Résineux et feuillus intolérants	Sapins et peupliers

S(E)	84	Le sapin et/ou l'épinette blanche occupent au moins 50% de la surface terrière de la partie résineuse du peuplement tandis que l'épinette noire et/ou rouge occupent au moins 25 % de cette même partie.	Résineux	Mélange de résineux	Sapins et épinettes
S(S)	84	Le sapin et/ou l'épinette blanche occupent au moins 75% de la surface terrière de la partie résineuse du peuplement.	Résineux	Un résineux dominant	Sapins
BbS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau blanc occupe plus de 50% de la surface terrière de la partie feuillue. Dans ce peuplement, le sapin constitue plus de 50% de la	Mélangés	Bouleaux et résineux	Bouleaux blancs et sapins
BjS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau jaune occupe plus de 50% de la surface terrière de la partie feuillue. Dans ce peuplement, le sapin constitue plus de 50% de la	Mélangés	Bouleaux et résineux	Bouleaux jaunes et sapins
FiS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le bouleau blanc et les peupliers occupent, en proportion à peu près égales, plus de 50% de la surface terrière de la partie feuillue. Dans ce	Mélangés	Feuillus intolérants et résineux	Feuillus intolérants et sapins
EE	92	Peuplement où les résineux représentent 75% et plus de la surface terrière totale et où l'épinette noire et/ou rouge occupent 75% et plus de celle de la partie résineuse. On donne alors au peuplement le nom de cette essence.	Résineux	Un résineux dominant	Épinettes
Epb	92	Peuplement d'épinette blanche généré par plantation	Résineux	Un résineux dominant	Épinettes blanches
ES	92	Peuplement où les résineux représentent 75% et plus de la surface terrière totale et où l'épinette noire et/ou rouge occupent de 50% à 74% de celle de la partie résineuse. Le reste de la partie terrière du peuplement est occupée par	Résineux	Mélange de résineux	Épinettes et sapins
F	92	Jeune peuplement dominé par les feuillus	Feuillus	Feuillus inclassables	Feuillus inclassables
M	92	Jeune peuplement composé d'un mélange de feuillus et de résineux	Mélangés	Mélangés inclassables	Mélangés inclassables
PeS	92	Peuplement mélangé où les feuillus représentent de 50% à 74% de la surface terrière totale. Le peuplier occupe plus de 50% de la surface terrière de la partie feuillue. Dans ce peuplement, le sapin constitue plus de 50% de la surface	Mélangés	Feuillus intolérants et résineux	Peupliers et sapins
R	92	Jeune peuplement dominé par les résineux	Résineux	Résineux inclassables	Résineux inclassables
SBb	92	Peuplement mélangé où les résineux représentent de 50% à 75% de la surface terrière totale. Dans ce peuplement, plus de 50% de la surface terrière de la partie résineuse est occupée par le sapin ou l'épinette blanche. Le bouleau	Sapins et bouleaux blancs	Mélangés	Résineux et bouleaux
SBj	92	Peuplement mélangé où les résineux représentent de 50% à 75% de la surface terrière totale. Dans ce peuplement, plus de 50% de la surface terrière de la partie résineuse est occupée par le sapin ou l'épinette blanche. Le bouleau	Sapins et bouleaux jaunes	Mélangés	Résineux et bouleaux
SS	92	Peuplement où les résineux représentent 75% et plus de la surface terrière totale et où le sapin occupe 75% et plus de celle de la partie résineuse. On donne alors au peuplement le nom de cette essence.	Sapins	Résineux	Un résineux dominant
SE	92	Peuplement où les résineux représentent 75% et plus de la surface terrière totale et où le sapin occupe de 50% à 75% de celle de la partie résineuse. Le reste de la partie terrière du peuplement est occupée par une ou plusieurs autres	Sapins et épinettes	Résineux	Mélange de résineux

Tableau 5A: Codes de hauteur (tiré de Rebout, 1998)

code hauteur	Inventaire	Descriptif	Classes hauteurs	Groupes dehauteur
2	73	La hauteur moyenne des dominants et des codominants se situe entre 17 m et 22 m	17-22 m	17-22 m
2	84	La hauteur moyenne des dominants et des codominants se situe entre 17 m et 22 m	17-22 m	17-22 m
2	92	La hauteur moyenne des dominants et des codominants se situe entre 17 m et 22 m	17-22 m	17-22 m
3	84	La hauteur moyenne des dominants et des codominants se situe entre 12 m et 17 m	12-17 m	12-17 m
3	73	La hauteur moyenne des dominants et des codominants se situe entre 12 m et 17 m	12-17 m	12-17 m
3	92	La hauteur moyenne des dominants et des codominants se situe entre 12 m et 17 m	12-17 m	12-17 m
4	73	La hauteur moyenne des dominants et des codominants se situe entre 7 m et 12 m	0-12 m	0-12 m
4	84	La hauteur moyenne des dominants et des codominants se situe entre 7 m et 12 m	7-12 m	0-12 m
4	92	La hauteur moyenne des dominants et des codominants se situe entre 7 m et 12 m	7-12 m	0-12 m
5	84	La hauteur moyenne des dominants et des codominants se situe entre 4 m et 7 m	4-7 m	0-12 m
5	92	La hauteur moyenne des dominants et des codominants se situe entre 4 m et 7 m	4-7 m	0-12 m
6	92	La hauteur moyenne des dominants et des codominants se situe entre 1,5 m et 4 m	1,5-4 m	0-12 m
6	84	La hauteur moyenne des dominants et des codominants se situe entre 1,5 m et 4 m	0-4 m	0-12 m

Tableau 6A: Codes de densité (tiré de Rebout, 1998)

Classes de pentes	code pente	Description
		Absence de données
0-3 %	A	Pente nulle
4-8 %	B	Pente faible
9-15 %	C	Pente douce
16-30 %	D	Pente modérée
31-40 %	E	Pente forte

Tableau 7A: Codes d'âge (tiré de Rebout, 1998)

code âge	Inven -taire	Descriptif	Âge	Groupes_âges	Classes_âges
me	73	Le peuplement mature ou suranné est régulier lorsque les arbres qui les composent ne forment qu'un seul étage.	mûr étagé	Prématuré, mature ou suranné	Étagé
mi	73	Le peuplement mature ou suranné est irrégulier lorsque la hauteur des tiges qui le composent présente une grande variation.	mûr irrégulier	Prématuré, mature ou suranné	Mature / suranné
mr	73	Le peuplement mûr ou suranné est étagé lorsque les tiges présentent deux étages bien distincts de couverture. L'étage inférieur est autre que celui de la régénération.	mûr régulier	Prématuré, mature ou suranné	Mature / suranné
r	73	Peuplement en régénération	en régénération	En régénération / jeune	En régénération
0	73	0 an	0 an	En régénération / jeune	En régénération
j	73	Jeune peuplement	jeune	En régénération / jeune	Jeune
0703	84	70 / 30 ans	70/30 ans	Étagé	Étagé
0100	84	1-20 ans	1-20 ans	En régénération / jeune	En régénération
0300	84	20-40 ans	20-40 ans	En régénération / jeune	Jeune
0307	84	30 / 70 ans	30/70 ans	Étagé	Étagé
0500	84	40-60 ans	40-60 ans	Prématuré, mature ou suranné	Prématuré
0700	84	60-80 ans	60-80 ans	Prématuré, mature ou suranné	Mature
0900	84	80-100 ans	80-100 ans	Prématuré, mature ou suranné	Suranné
30	92	Peuplement jeune	30-45 ans	En régénération / jeune	Prématuré
0	92	Peuplement coupé dans l'année.	0 an	En régénération / jeune	En régénération
15	92	Peuplement très jeune	15-30 ans	En régénération / jeune	Jeune
45	92	Peuplement prématuré	45-60 ans	Prématuré, mature ou suranné	Prématuré
60	92	Peuplement mûr	60-75 ans	Prématuré, mature ou suranné	Mature
75	92	Peuplement suranné	75-90 ans	Prématuré, mature ou suranné	Suranné
90	92	Peuplement suranné	90 ans et +	Prématuré, mature ou suranné	Suranné
1200	92	Cellule non forestière	Cellule non forestière		
1300	92	Cellule non forestière	Cellule non forestière		
1400	92	Cellule non forestière	Cellule non forestière		
1500	92	Cellule non forestière	Cellule non forestière		
1600	92	Cellule non forestière	Cellule non forestière		
1700	92	Cellule non forestière	Cellule non forestière		

Annexe B

Tableau 1B: Extrait des similarités entre les peuplements des inventaires 1984 et 1992 (calculées à partir du contexte de production)

P84/P92	3676	3110	3267	3075
1818	0.04388297872340	0.424933456151906	0.3366214126664	0.0438829787234
1817	0.04388297872340	0.252807856285086	0.1897384728691	0.1316489361702
1802	0.04388297872340	0.286495799547498	0.1751108132946	0.0438829787234
1816	0.30718085106382	0.0497936718879237	0.3223815852832	0.3949468085106
2358	0.26329787234042	0.2761524822695035	0.3793249726922	0.2632978723404
2395	0.04388297872340	0.2864957995474982	0.1751108132946	0.0438829787234
1812	0.30718085106382	0.2381801967106188	0.4135859905996	0.3071808510638
1803	0.30718085106382	0.2864957995474982	0.4135859905996	0.3071808510638
2357	0.0	0.2761524822695035	0.1408497953873	0.0
1813	0.04388297872340	0.3371674987050984	0.3764917323104	0.0438829787234
1811	0.07106694923945	0.1242617569943067	0.1762821514587	0.2069868018196
1815	0.04388297872340	0.2864957995474982	0.1751108132946	0.0438829787234
2394	0.0	0.2761524822695035	0.1408497953873	0.0
2355	0.0	0.3268241814271037	0.3422307144030	0.0
2393	0.0	0.2531727290114242	0.2037930577986	0.0
2356	0.0	0.2531727290114242	0.2037930577986	0.0
2354	0.0	0.2761524822695035	0.1408497953873	0.0
2353	0.0	0.2761524822695035	0.1408497953873	0.0
1810	0.04388297872340	0.2528078562850869	0.1897384728691	0.1316489361702
2352	0.0	0.2137223744014951	0.1789703627632	0.0
1814	0.04388297872340	0.2864957995474982	0.1751108132946	0.0438829787234
2351	0.26329787234042	0.3014883318483036	0.5548568875858	0.2632978723404
2392	0.26329787234042	0.3014883318483036	0.5548568875858	0.2632978723404
1804	0.30718085106382	0.2864957995474982	0.4135859905996	0.3071808510638
1819	0.30718085106382	0.2759972859391409	0.5003119078581	0.3071808510638
1808	0.30718085106382	0.336289471435422	0.4135859905996	0.3071808510638
1809	0.30718085106382	0.3371674987050984	0.6397896046508	0.3071808510638
1807	0.04388297872340	0.336289471435422	0.1751108132946	0.0438829787234
1805	0.04388297872340	0.2864957995474982	0.1751108132946	0.0438829787234
2349	0.26329787234042	0.2102348923237332	0.5342991108477	0.2632978723404
1806	0.30718085106382	0.2631511735630816	0.4135859905996	0.3949468085106
2345	0.0	0.1848990427449331	0.0954693236136	0.0
2348	0.26329787234042	0.2102348923237332	0.5342991108477	0.2632978723404
2331	0.26329787234042	0.1139184397163120	0.4053190058918	0.4264016954367
1801	0.30718085106382	0.3371674987050984	0.6397896046508	0.3071808510638
2342	0.0	0.2761524822695035	0.1408497953873	0.0
2343	0.01462765957446	0.0394503546099290	0.0248226950354	0.0877659574468
2350	0.0	0.2102348923237332	0.2710012385073	0.0
2346	0.26329787234042	0.2424645390070922	0.3793249726922	0.3510638297872
1800	0.30718085106382	0.2528078562850869	0.4135859905996	0.3949468085106

Tableau 2B: Extrait des similarités du modèle Matching Distance entre les peuplements des inventaires 1984 et 1992 (calculées à partir du contexte de production)

P84/P92	3676	3110	3267	3075
1818	0.04125	0.103125	0.04125	0.0
1817	0.04125	0.04125	0.04125	0.0
1802	0.04125	0.04125	0.04125	0.0
1816	0.103125	0.04125	0.103125	0.0
2358	0.2475	0.2475	0.2475	0.0
2395	0.04125	0.04125	0.04125	0.0
1812	0.28875	0.28875	0.28875	0.061875
1803	0.28875	0.28875	0.28875	0.0
2357	0.0	0.0	0.0	0.0
1813	0.04125	0.165	0.04125	0.0
1811	0.04125	0.04125	0.04125	0.0
1815	0.04125	0.04125	0.04125	0.0
2394	0.0	0.0	0.0	0.0
2355	0.0	0.12375	0.0	0.0
2393	0.0	0.061875	0.0	0.0
2356	0.0	0.061875	0.0	0.0
2354	0.0	0.0	0.0	0.0
2353	0.0	0.0	0.0	0.0
1810	0.04125	0.04125	0.04125	0.0
2352	0.0	0.061875	0.0	0.0
1814	0.04125	0.04125	0.04125	0.0
2351	0.2475	0.37124999999999997	0.2475	0.0
2392	0.2475	0.37124999999999997	0.2475	0.0
1804	0.28875	0.28875	0.28875	0.0
1819	0.28875	0.350625	0.28875	0.0
1808	0.28875	0.28875	0.28875	0.0
1809	0.28875	0.4125	0.28875	0.0
1807	0.04125	0.04125	0.04125	0.0
1805	0.04125	0.04125	0.04125	0.0
2349	0.2475	0.37124999999999997	0.2475	0.0
1806	0.28875	0.28875	0.28875	0.0
2345	0.0	0.0	0.0	0.0
2348	0.2475	0.37124999999999997	0.2475	0.0
2331	0.2475	0.2475	0.2475	0.0
1801	0.28875	0.4125	0.28875	0.0
2342	0.0	0.0	0.0	0.0
2343	0.061875	0.0	0.061875	0.0
2350	0.0	0.12375	0.0	0.0
2346	0.2475	0.2475	0.2475	0.0
1800	0.28875	0.28875	0.28875	0.0
2344	0.309375	0.2475	0.309375	0.0
1791	0.28875	0.4125	0.28875	0.0
1790	0.04125	0.165	0.04125	0.0

Tableau 3B: Extrait des similarités entre les peuplements des inventaires 1973 et 1992 (calculées à partir du contexte de bloc expérimental)

P73/P92	3676	3110	3267	3075
1	0.228662709084	0.14756802465	0.17734268515605	0.059114228385353
2	0.287776937469	0.50312358020	0.32534580243030	0.414669783940908
3	0.317406567099	0.44400935182	0.53289824071161	0.088743858014982
4	0.317406567099	0.44400935182	0.53289824071161	0.088743858014982
5	0.0	0.26666666666	0.26666666666666	0.0
6	0.177342685156	0.35512046293	0.44400935182272	0.0591142283853530
7	0.177342685156	0.35512046293	0.44400935182272	0.0591142283853530
8	0.19888804858	0.62178712960	0.44400935182272	0.0591142283853530
9	0.19888804858	0.53289824071	0.44400935182272	0.0591142283853530
10	0.0	0.0	0.0	0.0888888888888888
11	0.28777693746	0.41423469131	0.32534580243030	0.4146697839409086
12	0.406295455987	0.35497543205	0.44386432094881	0.4442994135705382
13	0.0	0.26666666666	0.26666666666666	0.0
14	0.228662709084	0.35512046293	0.44400935182272	0.0591142283853530
15	0.088453796267	0.26623157404	0.08845379626717	0.0294845987557234
16	0.088888888888	0.08888888888	0.08888888888888	0.2666666666666666
17	0.177342685156	0.17734268515	0.19888804858049	0.0854408451746460
18	0.444009351822	0.17734268515	0.28777693746937	0.1743297340635349
19	0.266231574044	0.53289824071	0.22851767821011	0.1150704748042757
20	0.266231574044	0.53289824071	0.22851767821011	0.1150704748042757
21	0.0	0.26666666666	0.0	0.0
22	0.177342685156	0.44400935182	0.14756802465252	0.0591142283853530
23	0.177342685156	0.44400935182	0.14756802465252	0.0591142283853530
24	0.177342685156	0.44400935182	0.19888804858049	0.0854408451746460
25	0.177342685156	0.44400935182	0.19888804858049	0.0854408451746460
26	0.0	0.0	0.0	0.0888888888888888
27	0.444009351822	0.17734268515	0.28777693746937	0.1743297340635349
28	0.532898240711	0.26623157404	0.31740656709900	0.2039593636931645
29	0.0	0.26666666666	0.0	0.0
30	0.177342685156	0.44400935182	0.19888804858049	0.0854408451746460
31	0.088453796267	0.08845379626	0.08845379626717	0.0294845987557234
32	0.266666666666	0.0	0.08888888888888	0.0888888888888888
34	0.177342685156	0.19888804858	0.17734268515605	0.1773426851560592
35	0.444009351822	0.55444360413	0.32534580243030	0.3253458024303012
36	0.266231574044	0.22851767821	0.53289824071161	0.5328982407116147
37	0.266231574044	0.22851767821	0.53289824071161	0.5328982407116147
38	0.0	0.0	0.26666666666666	0.266666666666666
39	0.177342685156	0.14756802465	0.44400935182272	0.4440093518227258
40	0.177342685156	0.14756802465	0.44400935182272	0.4440093518227258
41	0.177342685156	0.19888804858	0.44400935182272	0.4440093518227258
42	0.177342685156	0.19888804858	0.44400935182272	0.4440093518227258

Tableau 4B : Extrait des similarités entre les peuplements des inventaires 1984 et 1992 (calculées à partir du contexte de bloc expérimental)

P84/P92	3676	3110	3267	3075
1818	0.0	0.27222222222222225	0.13660194441531848	0.0
1817	0.0	0.27222222222222225	0.13660194441531848	0.0
1802	0.0	0.323542246150189	0.34162567204698896	0.0
1816	0.0	0.24788223418620559	0.2014167592301333	0.0
2358	0.0	0.24788223418620559	0.2014167592301333	0.0
2395	0.0	0.27222222222222225	0.13660194441531848	0.0
1812	0.0	0.27222222222222225	0.13660194441531848	0.0
1803	0.0	0.2089933452973167	0.17734268515605922	0.0
2357	0.0	0.18372087055632935	0.09259009004438898	0.0
1813	0.0	0.27222222222222225	0.13660194441531848	0.0
1811	0.0	0.20938088252031273	0.27036786782216676	0.0
1815	0.0	0.29788223418620563	0.14087861307598237	0.0
2394	0.0	0.20938088252031273	0.10619060750147813	0.0
2355	0.0	0.27222222222222225	0.13660194441531848	0.0
2393	0.0	0.4124311350390779	0.30017900936831665	0.0
2356	0.0	0.29788223418620563	0.14087861307598237	0.0
2354	0.0	0.29788223418620563	0.14087861307598237	0.0
2353	0.0	0.20938088252031273	0.10619060750147813	0.0
1810	0.0	0.29788223418620563	0.14087861307598237	0.0
2352	0.0	0.20938088252031273	0.10619060750147813	0.0
1814	0.0	0.03888888888888889	0.024074074074074078	0.08888888888888889
2351	0.0	0.4229067910566564	0.3352916791165488	0.044444444444444446
2392	0.0	0.24751269305461554	0.18652942897836539	0.13333333333333333
1804	0.0	0.2826978782398007	0.17171461416355055	0.0444444444444446
1819	0.0	0.2826978782398007	0.17171461416355055	0.0444444144444441
1808	0.0	0.33401790216776744	0.37673834179522103	0.044423444444444446
1809	0.0	0.2826978782398007	0.17171461416355055	0.044444444444444446
1807	0.0	0.24751269305461554	0.18652942897836539	0.13333333333333333
1805	0.0	0.2826978782398007	0.17171461416355055	0.044444444444444446
2349	0.0	0.3320624231462681	0.17171461416355055	0.044444444444444446
1806	0.0	0.2826978782398007	0.17171461416355055	0.044444424344446
2345	0.0	0.33401790216776744	0.37673834179522103	0.044444444444444446
2348	0.0	0.2826978782398007	0.17171461416355055	0.04411764444444446
2331	0.0	0.24751269305461554	0.18652942897836539	0.13333333333333333
1801	0.0	0.4229067910566564	0.3352916791165488	0.044444444444444446
2342	0.0	0.2826978782398007	0.17171461416355055	0.04444444444444446
2343	0.0	0.2455165505018746	0.3568005614983656	0.044444444
2350	0.0	0.24751269305461554	0.18652942897836539	0.13333333
2346	0.0	0.4229067910566564	0.3352916791165488	0.044444444
1800	0.0	0.33401790216776744	0.37673834179522103	0.044444444
2344	0.0	0.2455165505018746	0.3568005614983656	0.044444446

Tableau 5B : Extrait des similarités entre les peuplements des inventaires 1984 et 1992 (calculées à partir du contexte de production et du principe de variabilité)

P84/P92	3676	3110	3267
1818	0.049446202531645576	0.08564082278481014	0.049446202531645576
1817	0.049446202531645576	0.049446202531645576	0.049446202531645576
1802	0.049446202531645576	0.049446202531645576	0.049446202531645576
1816	0.08564082278481014	0.049446202531645576	0.08564082278481014
2358	0.29667721518987344	0.29667721518987344	0.29667721518987344
2395	0.049446202531645576	0.049446202531645576	0.049446202531645576
1812	0.346123417721519	0.346123417721519	0.346123417721519
1803	0.346123417721519	0.346123417721519	0.346123417721519
2357	0.0	0.0	0.0
1813	0.049446202531645576	0.1218354430379747	0.049446202531645576
1811	0.049446202531645576	0.049446202531645576	0.049446202531645576
1815	0.049446202531645576	0.049446202531645576	0.049446202531645576
2394	0.0	0.0	0.0
2355	0.0	0.07238924050632911	0.0
2393	0.0	0.036194620253164556	0.0
2356	0.0	0.036194620253164556	0.0
2354	0.0	0.0	0.0
2353	0.0	0.0	0.0
1810	0.049446202531645576	0.049446202531645576	0.049446202531645576
2352	0.0	0.036194620253164556	0.0
1814	0.049446202531645576	0.049446202531645576	0.049446202531645576
2351	0.29667721518987344	0.36906645569620256	0.29667721518987344
2392	0.29667721518987344	0.36906645569620256	0.29667721518987344
1804	0.346123417721519	0.346123417721519	0.346123417721519
1819	0.346123417721519	0.38231803797468356	0.346123417721519
1808	0.346123417721519	0.346123417721519	0.346123417721519
1809	0.346123417721519	0.4185126582278481	0.346123417721519
1807	0.049446202531645576	0.049446202531645576	0.049446202531645576
1805	0.049446202531645576	0.049446202531645576	0.049446202531645576
2349	0.29667721518987344	0.36906645569620256	0.29667721518987344
1806	0.346123417721519	0.346123417721519	0.346123417721519
2345	0.0	0.0	0.0
2348	0.29667721518987344	0.36906645569620256	0.29667721518987344
2331	0.29667721518987344	0.29667721518987344	0.29667721518987344
1801	0.346123417721519	0.4185126582278481	0.346123417721519
2342	0.0	0.0	0.0
2343	0.036194620253164556	0.0	0.036194620253164556
2350	0.0	0.07238924050632911	0.0
2346	0.29667721518987344	0.29667721518987344	0.29667721518987344
1800	0.346123417721519	0.346123417721519	0.346123417721519
2344	0.332871835443038	0.29667721518987344	0.332871835443038

Tableau 6B : Liste des liens sémantiques attendus au niveau détaillé (liens de référence au test de performance)

PEUPLEMENTS_1984	PEUPLEMENT_1992
2	9, 17, 26, 36, 40, 57,67
3	42, 46, 52,62
4	9, 15, 33, 55, 69,70
5	
6	33, 42, 55, 69,70
7	9, 15, 33, 55, 69,70
8	42
9	33, 42, 55,70
10	33, 55,70
11	3, 13, 22, 24, 28,45
12	5,11,14,19,21,23,48,54,56,58,59,60,61,65,68,72,75,76
13	33, 55, 69,70
14	33, 55, 69,70
15	3, 13, 22, 24, 28,45
16	42,7
17	13
18	9
19	9
20	42, 46, 52,62
21	13
22	9
23	3
24	3
25	9, 15, 33, 55, 69,70
26	7, 8, 10, 18, 27, 32, 51, 53,78

Tableau 7B : Liste des liens sémantiques attendus aux niveaux agrégés (liens de référence au test de performance)

COMPARTIMENTS	POLYGONES ÉCOLOGIQUES
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
COMPARTIMENTS	PEUPLEMENTS_92
1	9
1	17
1	26
1	36
1	40
1	57
1	67
2	46
2	52
2	62
3	9
3	15
3	69
3	70
4	33
5	3
5	13
5	22
5	24
5	28
5	45
7	55
8	42
9	13
10	9
11	7
11	8
11	10
11	18
11	27
11	32
11	51
11	53
11	78

POLYGONES ÉCOLOGIQUES	PEUPLEMENTS_84
1	2
2	3
2	20
2	8
3	4
3	7
3	25
4	6
4	9
4	10
5	11
5	15
5	23
5	24
6	12
7	13
7	14
8	16
9	17
9	21
10	19
10	22
11	26

Tableau 8B : Test Précision Rappel Modèle redéfini (niveau détaillé)

Rappel (%) Modèle **redéfini**	Précision (%) Modèle **redéfini**
2	100
3	66
7	70
13	65
14	74
15	34
17	28
20	30
21	31
22	33
25	29
30	35
33	38
43	43
53	48
55	42
56	43
58	44
59	32
63	19
65	21
70	17
71	18
73	19
75	15
76	15
77	9
78	10
83	9
86	10
90	8
91	9
94	9
95	8
98	7
99	6
100	5

Tableau 9B: Test Précision Rappel Modèle Matching Distance (niveau détaillé)

Rappel (%) Modèle **Matching Distance**	Précision (%) Modèle **Matching Distance**
0	0
0	0
0	0
0	0
0	0
0	0
0	0
2	50
2	50
2	50
6	28,57142857
6	28,57142857
6	28,57142857
8	30,76923077
8	30,76923077
8	30,76923077
8	30,76923077
8	30,76923077
8	26,66666667
15	18,51851852
15	18,51851852
15	18,51851852
15	17,44186047
15	17,44186047
15	17,44186047
15	14,42307692
15	14,42307692
15	11,81102362
15	11,81102362
19	14,07407407
19	13,66906475
50	34,96503497
67	44,07894737
67	44,07894737
76	28,04428044
81	21,25984252
86	13,45852895

Tableau 10B: Test Précision Rappel Modèle redéfini (niveau agrégé)

Rappel (%) Modèle **Redéfini**	Précision (%) Modèle **Redéfini**
5,88235294	100
5,88235294	100
5,88235294	100
5,88235294	100
5,88235294	100
5,88235294	100
5,88235294	100
5,88235294	100
5,88235294	50
11,7647059	50
23,5294118	66,6666667
29,4117647	55,5555556
29,4117647	55,5555556
47,0588235	57,1428571
47,0588235	44,4444444
47,0588235	34,7826087
70,5882353	41,3793103
82,3529412	45,1612903
82,3529412	45,1612903
94,1176471	43,2432432
100	37,7777778

Tableau 11B: Test Précision Rappel Modèle Matching Distance (niveau agrégé)

Rappel (%) Modèle **Matching Distance**	Précision (%) Modèle **Matching Distance**
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
5,88235294	16,6666667
5,88235294	16,6666667
5,88235294	16,6666667
5,88235294	16,6666667
5,88235294	16,6666667
11,7647059	33,3333333
11,7647059	14,2857143
29,4117647	18,5185185
41,1764706	18,9189189
41,1764706	18,4210526
58,8235294	20
64,7058824	19,6428571