Entrepôts de données dans le domaine spatial pour l'inventaire forestier

Marie-Dominique van Damme

Conservatoire National des Arts et Métiers

Cnam – Région Centre

Centre d'Enseignement Principal d'Orléans

MEMOIRE
Présenté par Marie-Dominique Van Damme
en vue d'obtenir

LE DIPLÔME D'INGENIEUR C.N.A.M.

en INFORMATIQUE

Entrepôts de données dans le domaine spatial pour l'inventaire forestier

Soutenu le 27 septembre 2010

Jury :

M. Fouad BADRAN, professeur CNAM Paris (Président)
M. Yves MARTINEZ, professeur CNAM Orléans, responsable de filière (Rapporteur)
Mme Laure KAHLEM, professeur CNAM Orléans
M. Nicolas POUSSIN, professeur CNAM Orléans
M. Jean-Luc Cousin, Chef du service informatique de l'Inventaire forestier national

Les travaux relatifs à ce mémoire ont été effectués à l'Inventaire Forestier National sous la direction de M. Jean-Luc Cousin.

Technologie sans simplicité n’est que gadget
HVD.

Remerciements

Je tiens d'abord à remercier Monsieur Claude Vidal pour m'avoir permis d'effectuer mon stage à l'Inventaire forestier national.

J'exprime toute ma gratitude à Jean-Luc Cousin pour sa contribution au bon déroulement de mon stage. Je lui adresse aussi mes remerciements pour l'indépendance qu'il a su me donner et de ses heures de relecture dominicales.

J'adresse également mes remerciements à Stéphanie Lucas pour ses multiples idées et ses diverses sources d'information instructives.

Enfin un grand merci à l'équipe des supers informaticiens toujours disponibles : Sylvain, Benoît, Rémi, ….

Je remercie également toute l'équipe du CNAM d'Orléans pour ces années où ils ont été disponibles et à l'écoute. Merci aux enseignants pour leur professionnalisme et pour la qualité de leurs cours. Tout particulièrement à M. Martinez qui a su me conseiller tout au long de cette formation.

Enfin, merci à vous, membres du jury qui avez accepté de juger ce travail.

Résumé

Dans l'open source ce n'est que depuis très récemment que des suites logicielles décisionnelles sont apparues. Demain des outils couplant les technologies BI aux données spatiales seront disponibles et semblent déjà très prometteurs.

Dans ce contexte, ce mémoire représente le résultat d'une étude technique dont le premier objectif est de prouver que la mise en place des entrepôts de données à l'Inventaire Forestier National dans le cadre d'une production de résultats statistiques sur tout ou partie des données est un projet réalisable tout en répondant à de nombreux besoins utilisateurs. Pour réaliser ce travail plusieurs outils permettant d'effectuer des agrégations sémantiques et spatiales de manière interactive, conviviale et rapide ont été développés.

Le second objectif du travail est de mettre en valeur les dimensions spatiales dans un processus de décision. Pour ce faire, un prototype d'application web cartographique a été construit, permettant une extraction interactive de connaissances géographiques par exploration dans les cubes de données.

Mots-clés: Analyse décisionnelle, Entrepôt de données, Modèle multidimensionnel, Cubes OLAP, Aide à la décision, Cartographie sur le web, Spatial OLAP.

Abstract

It is only very recently that decision-oriented softwares became available in Open Source, No doubt that tomorrow new and promising tools coupling BI technologies to spatial data sets will be available.

In this framework, this work is a technical study aiming primarily at demonstrating the feasibility of the implementation of datawarehouses at NFI (National Forest Inventory) allowing for the production of statistical results covering either extensively or in part the whole data set. In order to reach this objective, several interactive, fast, and user-friendly tools were developed, allowing for semantic and spatial aggregations.

A second objective of the work is to highlight the spatial dimensions in a decision-making process. A prototype of map-based web application is shown, allowing for the interactive extraction of geographical knowledge from exploration of the data cubes.

Keywords: BI, Datawarehouse, Multidimensional Model, OLAP, Decision support, Web mapping, Spatial OLAP.

Table des matières

Figures

Tableaux

Equations

Introduction

1 Contexte

La forêt française est une ressource naturelle importante par son étendue qui occupe 28,6% du territoire métropolitain et par l'enjeu de sa connaissance et de son état dans les nouvelles politiques économiques, sociales et environnementales tant au niveau national qu'international. L'Inventaire Forestier National se doit de répondre de façon adaptée à cette demande de la société. A cette fin, l'IFN a profondément modifié la plupart de ses services. Il est devenu un « observatoire » des espaces boisés et naturels français, doté d'outils modernes pour la connaissance et la décision.

Les systèmes d'informations décisionnels sont nés d'un besoin des entreprises à fournir aux décideurs des moyens d'accéder aux données de leurs propres systèmes dans le but de piloter leurs activités. Dès lors qu'il s'agit de faire du reporting ou de l'analyse de données pour arriver à fournir des tableaux de synthèse il fallait mettre en place des requêtes complexes, coûteuses en temps de réponse et en ressource informatique. Les structures des entrepôts de données, des bases de données multidimensionnelles et les outils d'exploration, par leur nature, sont construites pour supporter des analyses complexes et une découverte des connaissances. La fonctionnalité des OLAP est caractérisée par l'analyse multidimensionnelle et dynamique de données consolidées qui supportent les activités analytique et navigationnelle d'un utilisateur final. Les technologies de la Business Intelligence, comme les tableaux de bord, l'OLAP, le forage de données, sont disponibles commercialement depuis déjà une décennie, et l'open source s'attaque depuis quelques années à la gamme d'outils d'aide à la décision avec des solutions aujourd'hui très matures et en perpétuel progrès.

L'Inventaire forestier national possède un volume significatif de données complexes et a besoin de systèmes efficaces pour consulter et analyser les tendances pour la gestion durable des forêts et des disponibilités en bois. De plus ils doivent permettre d'accéder aux informations appropriées plus rapidement. C'est tout naturellement, que l'IFN s'intéresse à cette technologie.

L'IFN est en charge de cartographier la forêt française et exploite beaucoup de données spatiales. Elles ont un rôle fondateur dans le calcul et la diffusion des résultats statistiques. Les systèmes d'information décisionnels n'exploitent pas la composante géographique des données. La gestion et l'analyse des données géographiques sont le propre des systèmes d'information géographique (SIG). Les SIG reposent sur des processus transactionnels et ne sont pas capables de représenter et d'analyser les données géographiques intégrées dans des processus OLAP. C'est pourquoi une nouvelle technologie a émergé, connu sur le terme OLAP Spatial (SOLAP). Elle permet en particulier par la visualisation des résultats d'analyse sur une carte de comprendre la distribution géographique d'un phénomène et de permettre de le comparer à diverses granularités géographiques.

2 Les enjeux du projet

L'IFN réalise depuis novembre 2004 les opérations d'inventaire sur l'ensemble du territoire français au moyen d'un échantillon constitué pour une période de dix ans, dont un dixième est réactualisé chaque année. Au fil du temps, de nouveaux protocoles ont été mis en œuvre créant potentiellement une hétérogénéité sémantique et spatiale. A fréquence fixe, une nouvelle carte forestière est générée (le découpage des zones forestières évoluent, …). A chaque région peut correspondre un peuplement particulier (pin maritime des Landes, …). Occasionnellement une demande particulière doit être prise en compte (hauteur sans branche dans le Morvan, données sur les peupliers en période de crise économique, …). Chaque année le protocole peut évoluer ce qui se traduit notamment par des définitions de variables qui changent (cubage des arbres, détermination de l'âge moyen, ..).

Au sein de son système d'information, l'IFN dispose de quatre bases de données contenant les métadonnées, l'historique des opérations d'inventaire, les données collectées sur le terrain et la carte forestière. L'ensemble de ces données est traité par extraction et chargé dans une base d'exploitation annuellement. Ces bases sont actuellement utilisées pour fournir des résultats statistiques grâce à des applications métier. Ces applications utilisent un service de calcul pour accéder aux données.

A ce jour, l'IFN constate que les performances d'un tel système induisent des limites et que ces limites peuvent devenir contraignantes pour un utilisateur souhaitant obtenir de nombreux résultats en ligne pour une étude. Des outils plus adaptés et rapides amélioreraient la convivialité de ces applications et donc leur utilisation. La solution apportée par la problématique des entrepôts de données est dans les préoccupations de l'IFN, car celle-ci semble correspondre aux approches métiers et aux attentes des utilisateurs avec de meilleurs temps de réponses.

L'intégration des données géographiques dans l'analyse en ligne est aussi un enjeu majeur. La modélisation des entrepôts de données géographiques tout comme l'adaptation des fonctionnalités des systèmes d'entrepôt de données classiques pour les données géographiques est une problématique ouverte.

3 Contribution

Compte tenu des contraintes techniques liées à la spécificité du travail de l'IFN, le sujet de stage proposé correspond à une étude de faisabilité visant à mettre en place un entrepôt de données sur tout ou partie des données de l'IFN.

Intégrer les données de l'IFN dans un entrepôt de donnée constitue déjà un principal défi : il faut concevoir le modèle logique de l'entrepôt à partir du volume important des données acquises et disponibles issues de plusieurs sources et traitées avec des méthodes différentes, confronter les approches métier, et étudier le gain dans la prise de décision.

Les finalités de cet entrepôt sont présentées ici par ordre de priorité.

Le premier objectif est de produire des résultats statistiques. Ces résultats seront mis à disposition en utilisant des outils open source permettant d'effectuer des agrégations sémantiques et spatiales et offrant une grande interactivité. Une difficulté technique à prendre en considération dans la définition de l'entrepôt est l'intégration du calcul de la précision (variance) ou de la validité des résultats retournés (nombre de points : risque de seconde espèce). Ce calcul s'avére être complexe.

Un second objectif est de faire des analyses sur de gros volumes de données (globaux vers détaillés) sous forme de tableau de bord, de tableaux multidimensionnels, de graphes et de cartes, offrir une aide à la décision, déterminer des indicateurs de suivi et des analyses en temps réel. Il s'agit d'explorer les quelques sorties sur le décisionnel dans le monde de la gestion forestière.

Enfin en enrichissant la modélisation multidimensionnelle de manière à prendre en compte non seulement le caractère spatial ou géographique mais aussi le caractère multidimensionnel des données offrirait une nouvelle ouverture à l'établissement. Cependant la dimension spatiale au sein des entrepôts est une technologie émergente, prometteuse mais pas encore aboutie.

4 Plan du mémoire

Après ce chapitre introductif, nous structurons en cinq chapitres ce mémoire, présentant respectivement un état de l'art, une proposition en trois chapitres, et une conclusion.

L'état de l'art introduit les principaux concepts des technologies décisionnelles, leurs architectures ainsi que les outils d'aide à la décision. Les principes de la modélisation dimensionnelle

seront décrits car ils ont été utilisés dans la mise en place de l'entrepôt de données. Il vise aussi à dresser un panorama des caractéristiques principales de l'information géographique pour faire de la cartographie interactive sur le web. Enfin il présente un état des lieux des définitions de l'OLAP Spatial avec une présentation d'exemples de réalisation. Les produits open source prendront une part belle dans cet état de l'art.

La proposition est découpée en trois chapitres, dont le premier constitue un préambule aux deux suivants : il décrit de façon générale les besoins auxquels doivent répondre la mise en place d'un entrepôt de données : production de résultat, accès à des interfaces orientés navigation, fournir un cadre méthodologique, offrir aux utilisateurs des représentations géodécisionnelles. Il présente la méthode de calculs des résultats d'inventaire et une partie du système d'information de l'IFN dont l'objectif est de procurer ses résultats aux utilisateurs adaptés à leurs besoins. Cette présentation permettra au lecteur de mieux comprendre les choix technologiques et les détails d'implémentation des deux prochains chapitres.

Ensuite le chapitre 4 est entièrement dédié à la mise en place de l'entrepôt de données et le chapitre 5 à l'implémentation d'une interface Web OLAP mettant en valeur les dimensions spatiales. Ces deux chapitres présentent le travail réalisé et finissent par un bilan.

Enfin, la conclusion synthétisera le calendrier des réalisations, la démarche entreprise pour mener à bien cette étude, les bilans dressés au cours de l'étude, les réorientations choisies, la mise en production d'une partie du projet, ainsi qu'un petit bilan personnel.

État de l'Art.

1 Le décisionnel

Nous commencerons par décrire les systèmes décisionnels. Tout en présentant le vocabulaire communément utilisé dans les systèmes d'aide à la décision, nous décrirons les différents outils open source existants et utilisés dans la maquette géo-spatiale réalisée en fin de projet.

1.1 Aide à la décision

On appelle « aide à la décision » ou bien « décisionnel » ou encore « business intelligence », un ensemble de solutions informatiques permettant l'analyse des données de l'entreprise, afin d'en dégager des informations qualitatives nouvelles, de déceler des informations macroscopiques cachées dans de gros volumes de données. [SMILE]

1.1.1 Concepts principaux

Un entrepôt de données est une base de données utilisée spécifiquement dans le cadre de l'informatique décisionnelle. C'est un stockage intermédiaire des données de production dans lequel les utilisateurs finaux puisent avec des outils de restitution et d'analyse. Il permet à un analyste, qui manipule des données (très variées ou dans de gros volume de données) de prendre des décisions bonnes et rapides.

La finalité d'un entrepôt de données est de supporter le traitement analytique en-ligne (*OLAP*). Les techniques de type OLAP (*On-Line Analytical Processing*) effectuent la synthèse, l'analyse et la consolidation dynamique des techniques multidimensionnelles. Les techniques OLAP sont la manière la plus naturelle d'exploiter un entrepôt à cause de son organisation multidimensionnelle.

Les contraintes de performances et les fonctionnalités des applications OLAP ne sont pas les mêmes que les applications de traitement transactionnel en ligne (OLTP), habituellement supportés par les bases de données relationnelles. Les bases de données relationnelles sont conçues à partir de listes d'enregistrements. Les tables contiennent des informations qui peuvent être identifié de manière unique grâce aux champs clés. Les requêtes OLTP sont des parties de transaction. Elles concernent uniquement un petit nombre de tuples, mais apparaissent fréquemment. Au contraire les requêtes dans un entrepôt de données peuvent concerner des gigabytes de données, mais exécutées à une moindre fréquence. Attention ceci ne veut cependant pas dire que les sorties des entrepôts sont volumineuses ! En réalité les principes de structuration ne sont pas très nombreux et seront détaillés plus loin.

Pour permettre des analyses et des visualisations complexes, les données dans l'entrepôt sont organisées selon le modèle de données multidimensionnel. La modélisation de données multidimensionnelles (*modélisation dimensionnelle*) est le nom d'une méthode de conception logique souvent associée aux entrepôts de données. Elle signifie l'agrégation partielle des données de l'entrepôt selon différents critères. Un système OLAP emploie ce concept.

Quand on dit que l'information est multidimensionnelle, cela veut dire qu'elle peut-être représentée sous forme de tableaux croisés dynamiques. En effet elle est visible sous forme de cubes et les outils offrent des possibilités de naviguer dans ceux-ci en pivotant les axes, en consolidant les données à des niveaux hiérarchiques supérieurs, tout en désagrégeant d'autres données à des niveaux de détails très fins.

Dans le décisionnel on peut considérer deux modes de travail : le mode interactif et le mode rapport. Ils correspondent à des besoins différents mais cependant complémentaires :

- chercher une information à l'aveugle en effectuant différentes analyses, une restitution de données amenant à refaire une nouvelle recherche
- obtenir à partir d'une analyse prédéfinie une information quasi automatisée

Quelque soit ces situations il faut d'excellents temps de réponse pour que les utilisateurs puissent tâtonner en direct et une simplicité des outils pour que la configuration technique soit compréhensible par tous. L'expert garde toujours un rôle dans ce processus : dans le premier cas pour contrôler le résultat final produit et dans le second cas pour préparer le paramétrage réalisé en amont.

1.1.2 Architecture des systèmes décisionnels

Les architectures des systèmes décisionnels sont considérées comme des architectures à trois niveaux : l'entrepôt de données constitue le premier niveau, le service du deuxième niveau est instauré par le serveur OLAP et les clients sont mis en œuvre au dernier niveau, comme illustré par la figure 1 ci-dessous.

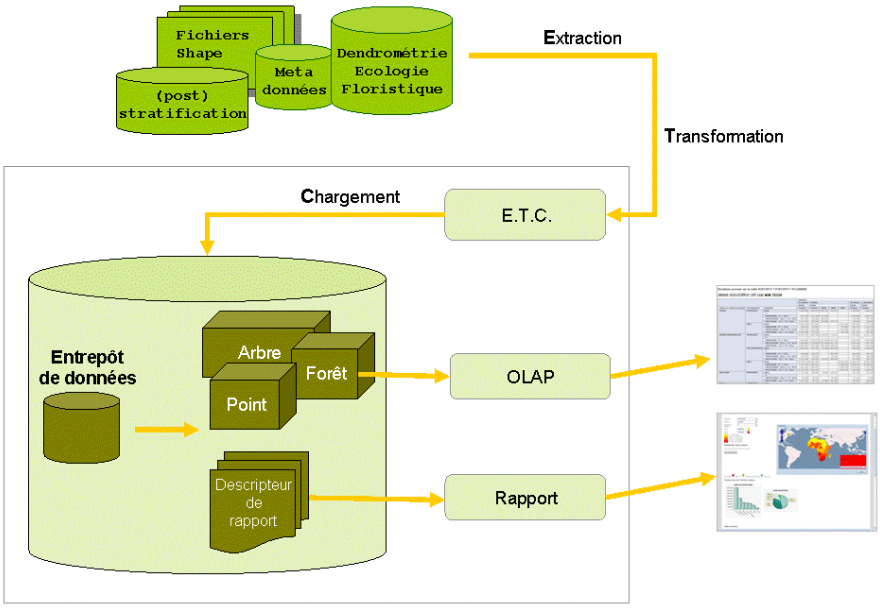


Figure 2-1. Architecture décisionnelle.

La finalité d'un entrepôt de données est de stocker et historiser des volumes importants de données. Les entrepôts de données sont alimentés grâce à des outils d'ETL (*Extract Transform and Load*) en français ETC (*Extraction, Transformation et Chargement*). Ces outils ont pour vocation d'extraire et de structurer des données en provenance des bases de données opérationnelles dites OLTP (*On Line Transactional Processiong*). Cette opération de rassemblement est difficile car les données sont hétérogènes, complexes et diffuses. La phase d'ETL réalise également un nettoyage des données suivi généralement d'une phase d'agrégation au sein des entrepôts.

A leur tour, ces données agrégées font l'objet d'une alimentation dans des bases de données multidimensionnelles appelées cubes OLAP. Le marché s'est structuré autour de deux approches : MOLAP et ROLAP.

L'approche MOLAP s'appuie sur une structure de stockage en cube, elle applique des techniques d'indexation et de hachage pour localiser les données lors de l'exécution des requêtes multidimensionnelles. Les temps de réponse sont faibles pour des calculs complexes, les systèmes

MOLAP fournissent une solution acceptable pour le stockage et l'analyse d'un entrepôt lorsque la quantité estimée pour les données d'un entrepôt ne dépasse pas quelques giga-octets et lorsque le modèle multidimensionnel ne change pas beaucoup [Guerrero].

Un outil ROLAP est capable de simuler le comportement d'un SGBD multidimensionnel autour d'un SGBD relationnel classique muni d'un moteur supplémentaire OLAP qui fournit une vision multidimensionnelle de l'entrepôt, des calculs de données dérivés et des agrégations à différents niveaux. Il génère les requêtes SQL mieux adaptés au schéma de l'entrepôt par une indexation spécifique et des vues matérialisées. Une vue matérialisée est la traduction relationnelle d'un cuboïde qui est pré-calculé et stocké dans l'entrepôt de données. Pour mesurer les performances d'un tel système, le facteur principal est la génération des requêtes SQL. Ces systèmes peuvent stocker de grands volumes de données mais peuvent aussi présenter des temps de réponse élevé.

Un cube est défini par un certain nombre de dimensions ou axes d'observation. Au croisement de ces dimensions se trouvent des mesures ou indicateurs. En général le cube permet des analyses ad hoc et des requêtes dynamiques ayant un caractère naturel et intuitif.

Les utilisateurs accèdent aux cubes OLAP grâce à des outils d'analyse offrant ainsi la capacité de réaliser à la volée des tableaux de synthèse, des rapports graphiques et des indicateurs pour réaliser des tableaux de bord.

Les opérations typiques exécutées par les clients OLAP peuvent être de nature :

- à relever le niveau d'agrégation (Roll-Up),
- à diminuer le niveau d'agrégation (Drill-down),
- de sélection,
- de projection (Slice et Dice),
- de réorienter la vue multidimensionnelle (Pivot).

Les technologies OLAP, par leur aspect dynamique et synthétique complètent les outils de reporting. Les outils de reporting sont généralement utilisés afin de fournir des vues statiques au travers de rapports instantanés à partir des données de l'entrepôt. A la différence des outils de requêtage OLAP, les fonctions de forage dynamique et de changement d'axes à la demande y sont absentes.

1.2 Modélisation multidimensionnelle

L'analyse multidimensionnelle est l'un des modes d'analyse le plus courants dans le décisionnel (les autres concernent les statistiques, le datamining, les systèmes d'aide à la décision, …).

Une base de données traditionnelle ne permet aux utilisateurs que des visions en deux dimensions comme par exemple l'étude des produits par région. Une base de données multidimensionnelle permet aux utilisateurs une analyse intégrant plusieurs dimensions comme par exemple, l'étude des ventes de produits par région par couleur, par taille et ce dans le temps.

1.2.1 Les dimensions (ou axes) et les hiérarchies

La structure de base de toute application multidimensionnelle est la dimension.

Une dimension est une liste complète d'éléments d'entrée (données qualitatives) et d'éléments calculés ou dérivés (à l'aide d'une formule quelconque).

La figure 2-2 illustre l'exemple de la dimension « Propriété d'un site » qui classifie le statut juridique d'une zone grâce à une hiérarchie sur deux niveaux (le premier niveau liste toutes les modalités de la classe propriété, le second niveau les regroupe en deux membres : « public » et « privé »).

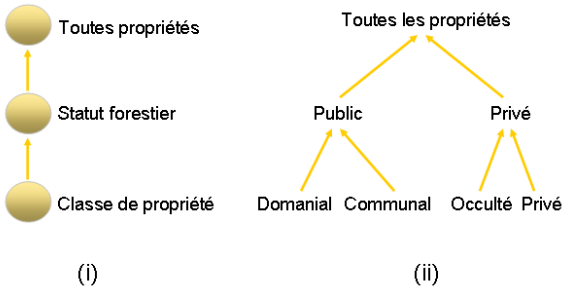


Figure 2-2. Dimension de la classe propriété. (i) Schéma (ii) Instance

Exemple IFN

Les dimensions sont organisées de façon hiérarchique. Chaque niveau d'une hiérarchie représente un degré de précision différent de l'information, on dit encore qu'elle représente une granularité différente de l'information.

Dans le cas d'une dimension hiérarchique, nous considérons la hiérarchie toute entière, et toutes les hiérarchies s'il y en a plusieurs, comme une seule et même dimension.

Quand dans une dimension on ne peut pas classer tous les niveaux sur une même hiérarchie on parle de hiérarchies multiples.

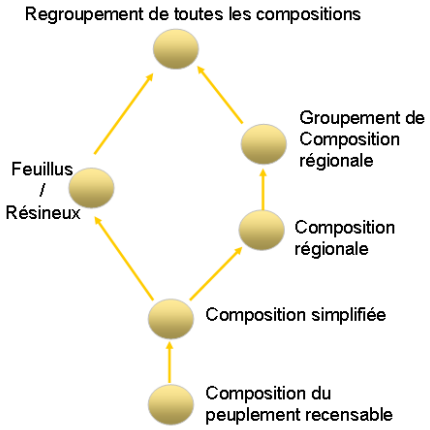


Figure 2-3. Dimension de la composition du peuplement recensable 2 hiérarchies

Le membre 'Feuillus' du niveau « Feuillus/Résineux » ne peut pas se décomposer par un membre 'Mélange mixte' du niveau « Composition régionale ».

Exemple IFN

On appelle dimension spatiale une dimension qui porte une propriété d'information spatiale (point, ligne, polygone, …).

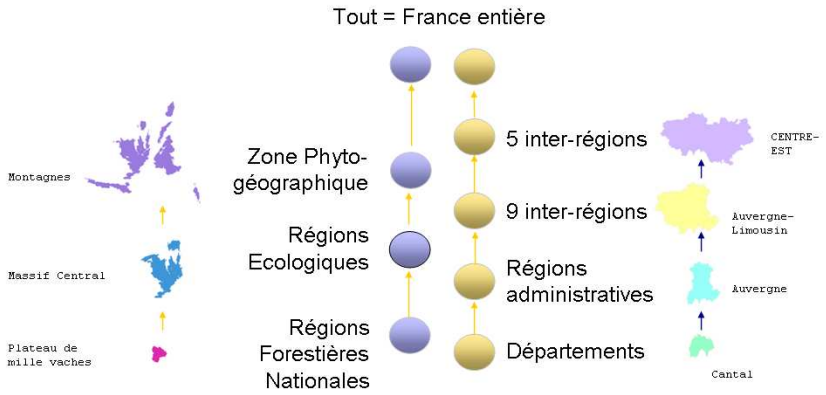


Figure 2-4. Différentes dimensions spatiales

A l'Inventaire Forestier National, les ressources forestières peuvent être vues spatialement suivant deux axes différents : par une subdivision administrative de la France ou par une subdivision écologique. Dans chacune de ces dimensions, chaque niveau a une propriété spatiale (son contour).

Exemple IFN

Chaque hiérarchie contient un niveau avec un seul membre : « Tout ». Cette notion est importante pour le calcul des agrégats. Les axes d'analyse fournissent grâce aux hiérarchies qu'ils portent les règles de calcul, ainsi que le les mécanismes de cheminement de la synthèse vers le détail.

1.2.2 Les faits et mesures

Les faits sont la clé de voute de la modélisation multidimensionnelle.

Un fait est un ensemble de mesures. Les mesures, qui sont aussi appelées variables ou métriques, fournissent une description quantitative du fait et identifie ce que l'on veut observer. Certaines mesures peuvent être calculées à partir d'autres mesures ou à partir des propriétés de membres d'une dimension.

Par exemple dans le domaine forestier une analyse multidimensionnelle sur le fait « ressource » permet de connaître la quantité de bois existant dans une zone à une date donnée en définissant comme mesure « volume sur pied de qualité 1», « volume sur pied de qualité 2 », « production nette en volume » et la mesure calculée « volume de bois d'œuvre » (qui est la somme des volumes de qualité 1 et 2).

Exemple IFN

Un fait se compose principalement de valeurs numériques (ce que l'on désire mesurer) identifiées par les clés associées aux dimensions. Le nombre des dimensions et leurs clés déterminent la finesse et la granularité des mesures. Dans les niveaux supérieurs, les mesures sont agrégées par les fonctions d'agrégations à partir des informations aux niveaux les plus détaillés.

Le volume d'un fait est très supérieur aux volumes des dimensions.

Ainsi dans l'exemple précédent, si le fait « ressource » est relié aux dimensions « Essence », « Structure forestière » et « Département », il permet de déterminer un volume sur pied ou une production nette par essence, par structure et pour chaque département. Le schéma multidimensionnel de la figure 2-5 s'associe à cet exemple. La production nette en volume pour une région est calculée par la somme des productions nettes en volume des départements de celle-ci.

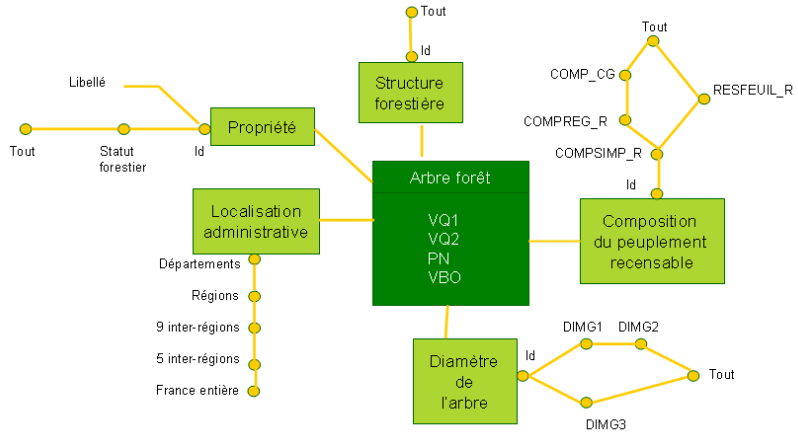


Figure 2-5. Schéma multidimensionnel "Ressource"

Exemple IFN

Lorsqu'un modèle dimensionnel se compose d'une table de faits centrale et qu'autour gravite un ensemble de tables plus petites nommées tables dimensionnelles, la structuration est dite un « schéma en étoile ». Cependant, une dimension peut être dans plus d'une seule table, fournissant un chemin pour joindre ces tables à la table de faits, ce type de schéma est dit en « flocons ». Le schéma en « constellation » est en fait composé de plusieurs schémas en étoile qui partagent des tables de dimension.

Les fonctions usuelles pour agréger sont les opérations classiques de SQL : la somme (« sum »), la moyenne (« avg »), le comptage (« count »), le minima (« min »), …. Dans les bases de données spatiales, il existe aussi des fonctions d'agrégats. Dans le tableau ci-dessous (tableau 2-1), nous montrons l'ensemble des fonctions d'agrégations pour les différents types de données (vu dans [SHEKHAR]).

	Fonctions d'agrégations		
Types de données	Fonctions distributives	Fonctions algébriques	Fonctions holistiques
Alphanumériques	Count, Min, Max, Sum	Average, MaxN, MinM	Median, MostFrequent, Rank
Spatiales : points, lignes, polygones.	BoundingBox, GeometricUnion, GeometricIntersection	Centroid, Center of gravity, Center of mass	Equi-partition, Nearest neighbor index.

Tableau 2-1. Opérations d'agrégations

Ces données numériques doivent être additives (somme sur toutes les dimensions) ou semi-additives (somme sur certaines dimensions). L'additivité est cruciale pour les outils décisionnels (l'utilisateur demande rarement l'analyse d'une seule ligne). Certains attributs ne sont pas additifs pour toutes les dimensions, par exemple un volume sur pied dans le temps ne s'additionne pas, mais il s'ajoute suivant toutes les autres dimensions.

Cette problématique dans OLAP est commune aux bases de données statistiques sous le nom « Summarizability ». Enfreindre cette condition donne des résultats faux, il faut respecter la sémantique des mesures.

1.3 Naviguer dans un Hypercube

La représentation physique d'un modèle multidimensionnel s'appelle un hypercube, on parle plus souvent de cube[1]. Dans l'espace multidimensionnel, les données des mesures correspondent aux données des cellules du cube qui seront analysées en fonction des données des dimensions qui constituent les axes d'analyse du cube. Les informations stockées correspondent dans les cellules aux valeurs des mesures détaillées et pour les axes les membres des niveaux les plus détaillés des différentes dimensions.

Un exemple d'hypercube est représenté dans la figure 2-6, il représente une coupe de l'implémentation physique du schéma multidimensionnel de la figure 2-5.

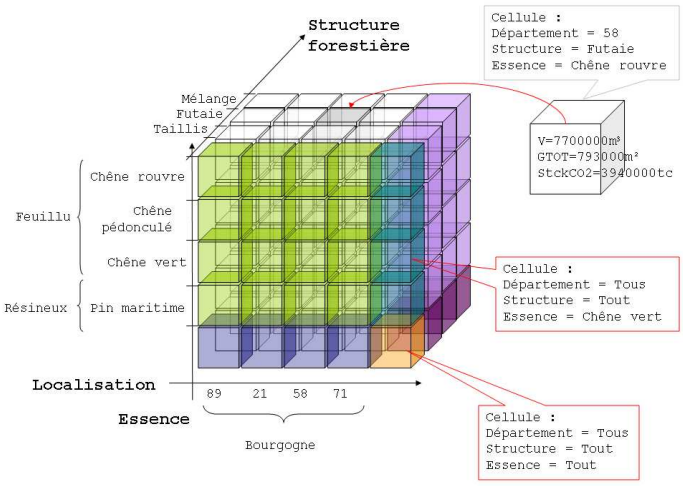


Figure 2-6. Hypercube

Exemple IFN

La navigation permet à l'analyste de visualiser les informations contenues dans le cube et de passer d'un niveau d'agrégat à un autre. C'est une exploration, l'utilisateur parcourt les données de l'hypercube selon les différents axes d'analyse à la recherche d'informations utiles, dans un processus fortement interactif, itératif et constructif, qui comprend des étapes de formulation des hypothèses, d'expérimentation et d'analyse. Ces chemins explorés sont imprédictibles.

Le processus interactif qui permet à l'utilisateur d'explorer un cube est géré par un client OLAP connecté au serveur OLAP. Les opérateurs les plus fréquemment utilisés sont :

- Les opérateurs de forage.
  - **Agrégation (ou Roll-Up)** : cette opération concerne le calcul pour une ou plusieurs dimensions. Elle permet de grimper dans les hiérarchies en agrégeant les mesures.
  - **Forer (ou Roll-down ou Drill-Down)** : le drill down est une opération de zoom avant. Il permet d'obtenir une information au niveau de détail plus fin en désagrégeant les mesures.

La structure hiérarchisée des dimensions permet une analyse en profondeur des données grâce à la technique du roll down et du roll up. Ces techniques permettent un forage progressif des

[1] En géométrie un cube désigne un solide à 6 faces carrées, dans le modèle multidimensionnel, un cube est une manière d'organiser des données, et en réalité ce n'est pas un cube. C'est une métaphore graphique afin de mieux comprendre le modèle.

données en passant par le niveau le plus élevé au niveau de détail le plus fin. Par exemple un propriétaire sera capable d'analyser les volumes par exploitabilité par région forestière, puis agréger rapidement les données au niveau zone écologique, puis au niveau national (drill up).

- Les opérateurs de coupe.

  - **Slicing :** un slice est une coupe ou un sous-ensemble d'un tableau multidimensionnel. Il permet de se concentrer sur une zone particulière d'un évènement.
  - **Dicing :** le dice est la sélection sur certaines valeurs de la dimension. Il permet de restreindre la dimension de l'hypercube.
  - **Rotate ou Pivot :** cette opération permet de changer l'orientation d'un cube, par exemple en inter-changeant les lignes et les colonnes du résultat.

  La structure en étoile permet une analyse selon divers axes grâce aux techniques de sélection, de projection, de réorientation des vues multidimensionnelles. Ces techniques permettent d'interagir itérativement pour reformuler et modifier les hypothèses. Par exemple l'expert forestier s'aperçoit que pour tous les peupliers dont le clone est à *null* (slice), en permutant cette colonne avec celle de l'homogénéité, l'indicateur de recensabilité est faux (rotate).

Les opérateurs OLAP ont une similitude avec les opérateurs des bases de données statistiques : les opérateurs de l'algèbre relationnelle : « select », « project », « union » et « aggregate ». Le tableau 2 ci-dessous, extrait de [MOR], résume cette correspondance :

BD Statistiques	OLAP
S-projection	Slice
S-sélection	Dice
S-aggrégation	Roll-Up
S-désagrégation	Drill-down
S-union	--

Tableau 2-2. Correspondance entre les opérateurs de BDS et OLAP

1.4 Outils décisionnels Open Source

Ce n'est que très récemment que les suites logicielles de solutions décisionnelles open source sont apparues. Elles se composent essentiellement de modules beaucoup plus anciens qui ont fait leurs preuves, et n'ont rien à envier aux solutions propriétaires. Il faut cependant faire attention aux produits pseudo open source, comme Pentaho, qui bride leur logiciel pour leur version gratuite.

Compte tenu des contraintes techniques liées aux spécificités du travail à l'IFN, le travail du stage consistant à une étude de faisabilité visant à mettre en place un entrepôt de données, il était préférable de s'attarder sur des modules dédiés uniquement à une tache spécifique de l'architecture décisionnelle, et non pas aux suites complètes qui auraient pénalisé en temps par leur complexité supplémentaire à implémenter et maintenir.

Ce n'est sûrement que partie remise, car leurs atouts (comme par exemple un générateur de rapport de tableaux multidimensionnels et de graphiques avec BIRT, le module cartographique pour SpagoBI) apporteraient une plus grande richesse et une plus grande ouverture au sujet de ce stage.

Pour ce travail, deux modules robustes des solutions décisionnelles open source ont essentiellement été mis en œuvre : Mondrian et JPivot. En ce qui concerne la base de données, le système PostgreSQL sera présenté dans la deuxième partie de ce chapitre, pour surtout détailler les fonctionnalités de la cartouche spatiale.

1.4.1 OLAP et Mondrian

Mondrian[2] est un serveur OLAP écrit en JAVA. Cet outil open source fait partie de la suite Pentaho Analysis Services, mais peut s'installer indépendamment. Mondrian utilise le langage d'interrogation MDX, le XMLA et les spécifications Java OLAP Interface (olap4j) afin de s'interfacer avec d'autres serveurs. C'est un serveur R-OLAP, c'est-à-dire qu'il lit par le SQL les sources de données et agrège le tout en mémoire cache.

Mondrian possède de nombreux atouts, il conserve de bonnes performances face à un grand volume de données en analyse interactive, il permet l'exploration dimensionnelle des données, il transcrit le MDX en SQL afin de construire ses requêtes dimensionnelles, il utilise les expressions calculées du langage MDX afin de produire des calculs avancés.

Comment fonctionne le serveur pour l'analyse du calcul et l'agrégation des données ? Sa couche dimensionnelle commence par décomposer, valider et exécuter les requêtes MDX. Une requête est évaluée en plusieurs phases : les axes d'analyse sont les premiers à être calculés, puis les valeurs des cellules sont traitées à travers l'indication de ces axes, tâche dédiée à la couche suivante.

La couche « étoile » est responsable du traitement des données agrégées stockées dans le cache. La couche dimensionnelle envoie des requêtes concernant un ensemble de cellules. Si cet ensemble n'est pas dans le cache, elle va faire appel à la couche de stockage. Un optimiseur de requête autorise la manipulation des requêtes existantes dans le cache afin d'améliorer les performances du serveur, plutôt que de partir de rien.

La couche de stockage est le système de gestion des bases de données. Son rôle est de produire des données agrégées par rapport aux dimensions à partir des données stockées dans l'entrepôt de données. A l'IFN, la plupart des données sont stockées dans des bases PostgreSQL, ce que nous continuerons à utiliser pour l'entrepôt de données.

Comment définit-on un cube à Mondrian ? Le modèle logique des cubes sont représentés à l'aide d'un document XML. Il suffit d'en éditer un et de respecter la syntaxe de Mondrian afin de réaliser une base de données multidimensionnelle. Les composants les plus importants dans le cube sont les cubes, les dimensions et les mesures. En annexe 1, un extrait du schéma de la constellation de l'entrepôt des données dendrométriques et écologiques de l'IFN est présenté.

Multi-Dimensional Expression (MDX)

MDX est le langage principal implémenté par Mondrian. Créé en 1997 par Microsoft, ce langage est de plus en plus utilisé. Une requête ressemble à ceci :

```
Select
{
  ([Measures].[Nombre de points Forêt], [Essence principale].[total].[CONIFERE]),
  ([Measures].[Surface Forêt], [Essence principale].[total].[FEUILLU])
} ON COLUMNS,
{
  ([Age calculé].[total].Children)
} on rows
From [Foret]
Where ([Momentanément déboisé].[total].[POINT BOISE])
```

A première vue, MDX semble assez proche de SQL, mais est différent. En reprenant l'image du cube, chaque cellule a un nom, et ce nom peut se définir à partir d'une autre cellule. Par exemple :

[ANNEE.2007] = [ANNEE.2008.PrevMember].

Ce langage permet de naviguer dans les hiérarchies. Les membres calculés sont une propriété assez puissante de l'analyse OLAP et permettent de créer des mesures dont les valeurs viennent non

[2] Site web : http://mondrian.pentaho.org

pas de colonne des tables de fait, mais d'autres formules MDX. Dans le chapitre 4, les membres calculés utilisés dans le cadre de l'entrepôt de données sont expliqués plus en détail.

1.4.2 Client OLAP : JPivot

JPivot est un client OLAP (API Java) disposant d'une interface web qui permet aux utilisateurs d'exécuter de manière très interactive et conviviale des navigations dans une base de données dimensionnelles et d'en extraire des tableaux croisés. Couplé à Mondrian, ce duo logiciel performant est employé dans la plupart des suites décisionnelles open source.

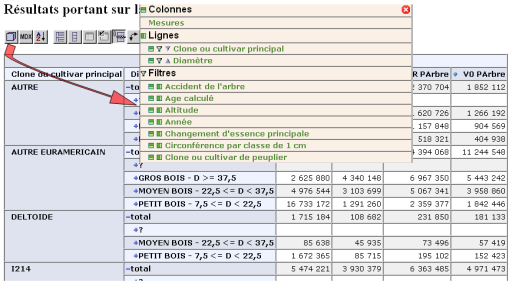


Figure 2-7. Tableau extrait de JPivot

Le module de génération automatique de graphiques n'est pas très exploitable pour des publications, il permet de déceler des tendances au fur et à mesure. A l'inverse le module extracteur Excel qui permet d'exporter des tableaux afin de réutiliser les résultats produits, fonctionne très bien et avec des options pratiques : choix des valeurs nulles, regroupement des cellules, ….

La facilité informatique ne doit cependant pas faire oublier à l'utilisateur à ne remonter de l'entrepôt de données que les résultats utiles, non pas pour des problèmes de performance du serveur Mondrian, mais tout simplement parce que les interfaces utilisateurs (navigateurs) deviennent très vite inefficaces lorsqu'il faut charger des tableaux croisés comportant trop de lignes et de colonnes.

2 Cartographie sur le web

Dans le paragraphe précédent, nous avons parlé de dimension géographique qui peut restituer des informations suivant des propriétés spatiales, comme par exemple le contour d'un territoire. La restitution géographique est un véritable atout. Aussi nous décrirons les caractéristiques principales de l'information géographique : comment la stocker, l'afficher, la restituer, …. Les systèmes d'information géographique regroupe des outils d'acquisition de données géographiques, mais intègre aussi des fonctionnalités d'assemblage, d'analyse et de visualisation. Ce sont ces capacités que nous décrivons ci-dessous.

2.1 Bases de données spatiales

Une base de données spatiale est une base de données dans laquelle les informations peuvent être localisées géographiquement, c'est-à-dire des données relatives aux objets dans l'espace y compris les points, les lignes et les polygones. Malgré leurs spécificités, les SGBD spatiaux peuvent être vus comme des extensions des SGBD relationnels.

	RELATIONNEL	**SPATIAL**
Données	Entier, Réel, Texte	Plus complexes : Point, Ligne, Région, …
Prédicats et calculs	Tests : ≤, =, … Calculs : +, *, … Fonctions simples	Prédicats et calculs géométriques et topologiques Tests : intersecte, adjacent à, …. Fonctions géométriques : intersect, surface, ….

Manipulation	Opérateur de l'algèbre : Sélection, Projection, Jointure, … Agrégats : count, sum,, avg, …	Manipulation par thème ou inter-thème Sélection et jointure sur critères spatials Agrégats : fusion d'objets adjacents.
Liens entre objets	Par clés de jointures	Liens spatiaux (souvent) implicites.
Méthodes d'accès	Index B-tree, hachage	Index R-Tree, quad-tree, etc.

Tableau 2-3. Comparatif des SGBDR et des SGBDS

Du coté opératoire, les simples prédicats de comparaison entre les valeurs alphanumériques dans un langage de requête comme SQL ne suffisent plus. Ils trouvent leurs équivalents en des opérateurs et prédicats pour capturer les relations spatiales pouvant exister entre les objets spatiaux. Ils sont basés sur des algorithmes géométriques implémentés dans les SGBD spatiaux.

Du point de vue du langage de requête SQL, tous les opérateurs de bases comme la sélection, la jointure trouvent un équivalent spatial. Plus particulièrement, pour l'opérateur de jointure, les algorithmes trouvent leurs limites et ont été remplacés par de nouveaux algorithmes spécifiques.
Ces développements montre qu'n SIG n'est pas une simple extension d'un SGBDR.

Rappelons tout d'abord les principaux avantages des bases de données spatiales : elles sont rapides (comparées aux traitements des fichiers Shape), elles sont optimisées pour de large volume (index spatiaux), permettent un accès simultané et avec sécurité à plusieurs utilisateurs et possèdent une multitude de fonctions afin d'accéder facilement aux données.

Les index spatiaux rendent possible l'utilisation de base de données spatiales avec de grandes quantités de données. L'objectif des méthodes d'accès par index spatiaux est l'accélération de la recherche d'objets géographiques, de limiter les comparaisons géométriques et organisé les données par proximité géométrique de manière à minimiser les entrées-sorties. Deux méthodes d'indexation sont proposées :

- Indexation à l'aide de Quadtree : c'est un découpage successif de l'espace en carrés (pavage, tesselation).
- Indexation à l'aide de R-Tree (par défaut) : c'est une utilisation des rectangles minimum englobants.

On peut indexer une géométrie avec soit Quadtree, soit R-Tree, soit les deux.

Dans le monde des bases de données spatiales, PostgreSQL et sa cartouche PostGIS[3] sont considérées comme un pilier dans la communauté open source GIS. L'IFN essaie d'opter pour un maximum de logiciels open source, c'est tout naturellement que l'établissement a choisi d'implémenter ce SGBD depuis de nombreuses années déjà (2003).

Les informations spatiales d'un objet géométrique sont stockées dans un nouveau type de donnée : *geometry*. En créant une nouvelle colonne avec ce type, le SGBD ajoute simultanément un *srid*, code ESPG[4] qui contient l'identifiant de la référence spatiale, plus précisément le système de coordonnées géoréférencées de projection de la donnée (la table *spatial_ref_sys* référence la liste des codes EPSG que PostGIS utilise).

Par exemple, ajoutons en ligne de commande un champ spatial, un contour géographique, à une table REGN, table qui contient déjà les informations alphanumériques des régions forestières nationales. La syntaxe de base de cette commande est la suivante :

```
AddGeometryColumn (<table name>, <column name>, <srid>, <datatype>, <numdimensions>)
```

[3] http://postgis.refractions.net/
[4] Code utilisé dans les standards de l'OGC (Open Geospatial Consortium)

Quelle projection prendre ? En choisissant « -1 » comme SRID, on élude tout choix. Aucune opération de changement de coordonnées n'est prévue dans le développement de l'entrepôt de données et dans l'implémentation de la maquette, nous aurons juste à nous assurer que toutes les informations spatiales sont saisies avec le même système de projection. Aujourd'hui, l'Inventaire Forestier emploie le système de projection Lambert 2 étendu (SRID = *27582*) et Lambert 93 (SRID = *2154*).

PostGIS supporte les principaux types de données spatiales : POINT, LINESTRING, POLYGON, MULTIPOINT, MULTILINESTRING, MULTIPOLYGON et GEOMETRYCOLLECTION.

Le cinquième paramètre décrit le nombre de dimension de l'élément géographique : si l'objet est un élément du plan, sa dimension est 2, dans l'espace un point nécessite trois coordonnées (x,y,z), sa dimension est alors 3.

Pour notre exemple, la ligne de commande devient :

```
AddGeometryColumn ('regn','regn_geom', '2154', 'MULTIPOLYGON', 2);
```

Une fois le champ spatial mis en place, voyons comment insérer de nouvelles valeurs dedans. Plusieurs solutions sont possibles : on connait la représentation textuelle du contour, on dispose d'un shapefile contenant sa description, on dispose d'informations spatiales annexes.

Pour le premier cas, on dispose de la requête traditionnelle insert. Pour ajouter le contour géographique du département du Loiret en coordonnées Lambert 2 étendue dans une table :

```
INSERT INTO departement (dep, the_geom) VALUES ('45', GeomFromText('POLYGON((361 201,362
211,382 217,407 225,406 215,413 212,413 196,397 197,386 193,384 189,381 189,378 196,371
202,361 201))',-1));
```

La manière la plus simple est d'importer un fichier shape contenant les régions forestières nationales, téléchargé par exemple sur le site de l'IFN[5], en employant le script *shp2pgsql*. Cette commande introspecte le fichier, crée la table en ajoutant le champ spatial correspondant, parcourt chaque ligne et crée les requêtes insert correspondantes.

```
shp2pgsql –s <SRID> <SHAPEFILE> <TABLENAME>
shp2pgsql –s -1  rnifn250.shp regn
```

Une troisième méthode consiste à manipuler des données spatiales déjà existantes en base. Par exemple, afin d'ajouter les régions administratives françaises, il suffit d'utiliser la requête ST_UNION :

```
insert into regionadm (ra, ra_libelle, ra_geom)
select '10', 'Région Centre', st_union(dep_geom)
from departement
where dep in ('18', '45', '41', '36', '28', '37')
```

PostGIS, comme toutes les bases de données, est pourvu d'index afin d'augmenter ses performances sur de grandes tables. L'index spatial *GIST* (Generalized Search Tress) permet d'augmenter les performances des requêtes spatiales. Traditionnellement, il référence une clé de type entier qui permet de court-circuiter la recherche au lieu de scanner toutes les lignes de la table. Pour les index spatiaux l'idée reste la même : on crée un index sur la bouding box de la géométrie du champ que l'on interroge souvent (appelé aussi MBR : *minimum bounding rectangle*). PostgreSQL supporte

[5] Disponible sur le site de l'IFN : http://www.ifn.fr/spip/article_special.php3?id_article=134

les index de type R-Tree, mais leur implémentation est jugée [CNES] moins puissante que celle de type GIST. De plus le type d'index GIST offre deux avantages : il permet l'indexation de colonnes contenant les valeurs nulles et il permet d'indexer de gros objets en ne prenant en compte que leur MBR, ce que ne ferait pas un index R-Tree, limitant l'indexation des objets de moins de 8Ko.

Dans le paragraphe suivant, nous verrons comment visualiser ces données dans un navigateur, cependant les données spatiales peuvent être consultées à partir d'applications bureau. Dans le monde de l'open source, nous pouvons citer : QGis[6] (QuantumGIS). Valable sur les systèmes Unix, Linux, Windows et MacOS, il peut lire des données spatiales sous PostgreGIS et accepte une foule de formats d'entrée/sortie (Vecteur : SVG, Matriciel : JPEG, TIFF, PNG, des fichiers SIG : MIF/MID, SHP, et se connecte à des services web en WMS), GRASS, etc.

2.2 Serveurs et clients OGC

Après avoir décrit les supports des données spatiales nous allons présenter la famille des OGC Web Services. Ils permettent, entre autres, de télécharger des données vecteurs, d'afficher des cartes dans votre navigateur préféré, …. Nous parlerons de la mise en œuvre de MapServer, des SLD, d'OpenLayers et de GeoExt, logiciels open source utilisés au cours de cette étude.

2.2.1 OGC Web Services

L'OGC (Open Geospatial Consortium) est une collaboration de différents acteurs internationnaux dédiée au développement des systèmes ouverts de l'information géographique et dont l'objectif est de garantir ces systèmes cartographiques interropérables dans leurs contenus, leurs services, leurs échanges. Son travail aboutit à des spécifications de langages, de formats ouverts, il coordonne les applications open source dans le domaine de la géomatique.
Parmi les recommendations faites par l'OGC, nous ne présenterons que les deux les plus implémentés : le WMS et le WFS.

- **WMS** ou **Service Web de Cartographie**. Il normalise la manière selon laquelle des applications clientes doivent demander une carte (sous une image ou sous un format vecteur) de données géoréférencées à partir de différents serveurs et la manière selon laquelle ils doivent décrire les données qu'ils sont capables de fournir [georezo]. Le client peut demander au serveur implémentant le service WMS les trois opérations suivantes :

1) GetCapabilities : cette requête retourne les services du serveur OGC, comme les méta-données qui décrivent le contenu du service et les paramètres acceptés. Le format retour est un fichier xml.

2) GetMap : une fois que vous connaissez quels sont les layers disponibles via le getCapabilities, cette requête vous retourne une image d'une carte. Le client spécifie les dimensions, le format, la méthode de projection, … dans l'url de la requête.

Exemple d'URL :

```
http://localhost:8888/cgi-
bin/mapserv.exe/?map=D:/localisation.map&SRS=epsg=27572&EXCEPTIONS=application%2Fvnd.ogc.s
e_inimage&bbox= 47860, 1197822, 1620431, 2677441&LAYERS=dim_dep&FORMAT=png
```

L'option SRS définit le système de référence spatiale utilisé par l'OGC, c'est un code identifiant le système coordonnées pour la projection utilisée; 27572 correspond à du Lambert 2 étendu.

[6] Site de Qgis : http://qgis.org

3) GetFeatureInfo : retourne des informations sur les objets représentés sur la carte.

Un service WMS n'est pas habituellement appelé à travers un navigateur, il est le plus souvent appelé via une application cliente qui fournit des options afin de configurer les options de manière interactive.

- **WFS** ou **Service Web de Fonctionnalités**. Il interroge via une URL des serveurs cartographiques afin de manipuler des données géographiques. La sortie de la requête est prête pour l'affichage, toutes les opérations se déroulent sur le serveur. La spécification permet trois niveaux de fonctionnalités : la première opération consiste à retourner les métadonnées, la deuxième permet l'extraction de base des données et la troisième opération offre la possibilité de les manipuler.

1) GetCapabilities : retourne, comme la requête du service WMS, les métadonnées décrivant les types de données disponibles sur le serveur. Le format du fichier xml de retour est légèrement différent.

2) DescribeFeatureType : permet de retourner la structure des types d'entités (multipolygone qui a une surface et un périmètre comme propriétés) retournés par la requête précédente. Le format de retour est un fichier xml.

3) GetFeature : retourne les entités (géométrie et/ou attributs) en GML (Geography Markup Language), langage standard ouvert mais très verbeux (le flux peut être compressé au retour).

4) Transaction : permet de créer l'entité, le mettre à jour ou le supprimer.

5) LockFeature : permet de bloquer des objets lors d'une transaction.

Les spécifications visant un service WFS précise que les entités doivent être échangées au moyen du format GML.

- **SLD** ou **Description du Dtyle des Couches** : un standard de l'OGC qui ajoute de nouvelles fonctionnalités aux standards établis comme le WMS et le WFS.

Ce mécanisme permet de modifier la mise en forme des données géographiques, par conséquent l'aspect de la carte. Le but est de séparer complètement le style de la donnée, comme les feuilles de style CSS pour les pages HTML. Par exemple, il permet à l'aide de critères définis, de modifier la couleur des départements suivant la valeur du champ « population qui a voté ».

Un SLD est un fichier XML qui décrit son numéro de version, son nom, le style à appliquer. Il suffit de spécifier l'URL où est stocké le fichier XML pour que celui-ci soit appliqué aux flux de données. Par exemple :

```
<Rule>
  <Name>MPYR</Name>
  <ogc:Filter>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>code</ogc:PropertyName>
      <ogc:Literal>MPYR</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:Filter>
  <PolygonSymbolizer>
    <Fill>
      <CssParameter name="fill">#4169E1</CssParameter>
    </Fill>
  </PolygonSymbolizer>
</Rule>
```

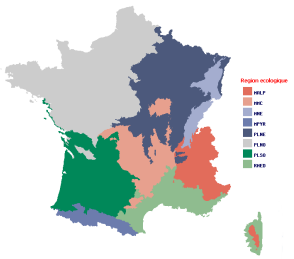


Figure 2-8. Représentation des grandes régions écologiques

L'architecture du système d'information de l'IFN, plus particulièrement la chaîne de traitement et de diffusion d'information, repose sur une architecture orientée Web Services (SOA). L'ensemble des applications cartographiques de publication web (découverte des données brutes, application d'organisation du travail, l'accès aux données de la carte forestière, …) repose sur des logiciels GIS open source de l'OGC : plus particulièrement sur le serveur cartographique MapServer. Aussi, pour la réalisation de la maquette de l'application décisionnelle web cartographique, il était légitime de choisir ce même environnement.

Exemple IFN

MapServer :

MapServer[7] est un serveur open source mature de l'OGC. Il se déploie directement sous Apache ou en mode CGI. Il permet la réalisation d'applications web à composantes cartographiques et beaucoup d'applications gravitant autour du serveur comportent d'outils puissants développés en PHP. Il permet d'afficher des données géographiques sous forme vectorielle ou cartographique. MapServer est multiplateforme.

Les MapFiles constituent le mécanisme de base : il contient la localisation des sources de données, les références, le format de sortie, les projections, …. C'est le fichier de configuration du serveur. Il est composé d'une hiérarchie d'objet, dont les objets LAYER qui enregistrent les couches de données de l'application.

Comment ajouter un layer Postgis ?

```
LAYER
STATUS ON
NAME layer_name
TYPE POLYGON
CONNECTIONTYPE postgis
CONNECTION 'user=USER password=PASS host=HOST dbname=DBNAME'
DATA 'the_geom FROM (SELECT the_geom, gid from departement)
     as foo using unique gid using srid=-1'
CLASS
STYLE
COLOR 150 80 60
OUTLINECOLOR 30 30 30
END
END
END
```

(i)

Comment ajouter un layer de type vectoriel ?

```
LAYER
STATUS ON
NAME layer_name
TYPE POLYGON
DATA 'path/data.shp'
CLASS
STYLE
COLOR 180 180 240
OUTLINECOLOR 0 0 180
END
END
END
```

(ii)

Les données de type ShapeFile sont les plus simples à utiliser, (ii) en illustre un exemple de configuration. MapServer dispose aussi de son propre driver Postgis, (i) illustre un exemple de configuration d'une couche base de données.

A partir de la version 5.x, MapServer permet de réaliser de la géostatistique intégrant un module qui génère des histogrammes et des graphiques.

Nous venons de voir comment les serveurs OGC géraient les couches de données, les options de style ou les codes EPSG autorisés, tout ce dont les applications clientes n'auraient plus à s'occuper. Une requête WMS GetMap permet de spécifier tous les derniers détails afin de restituer la carte demandée.

Regardons maintenant les rôles des applications et des librairies web qui permettront de valoriser ces données cartographiques.

[7] Site de MapServer : http://mapserver.gis.umn.edu/

2.2.2 OGC Clients

Ce paragraphe permet de vous présenter les applications clientes qui utilisent les services WMS et WFS. Nous commencerons par OpenLayers, Framework javascript et terminerons par GeoExt, librairie un peu plus puissante.

- **OpenLayers**[8] : est une bibliothèque Ajax open source permettant la mise en place, de façon dynamique, d'interfaces cartographiques dans des pages web.

L'intégration d'OpenLayers s'opère très facilement. Après avoir placé le répertoire des sources dans l'environnement de développement web, il faut ensuite ajouter à l'intérieur du header de la page web la localisation du répertoire de la manière suivante :

```
<script src="./js/OpenLayers/OpenLayers.js" type="text/javascript">
```

Afin d'intégrer une carte, il faut créer un objet map (1) qui prend comme paramètre l'identifiant de la balise qui contiendra la carte, et qui permettra de la manipuler. Ensuite il faut créer un layer (2), objet qui va interroger un serveur cartographique WMS. Enfin il reste juste à afficher la carte désirée (3).
Avec ce code, on obtient la carte des grandes régions écologiques françaises vue dans la figure 1 un peu plus haut.

```
<script type="text/javascript">
function init() {
    var map, ol_wms;
    // (1) instanciation du constructeur
    map = new OpenLayers.Map('mapREco');
    // (2) Choix de la couche des régions écologiques
    var re_wms = new OpenLayers.Layer.WMS(
              "Régions écologiques",
              "http://localhost:8888/cgi-bin/mapserv.exe/?map=C:/localisation.map", {
              layers: "zpg0",
              format: "png"
    });
    // (3) ajout des couches a la carte
    map.addLayers([re_wms]);
    //On zoom au max
    map.zoomToMaxExtent();;
}
</script>
```

OpenLayers est attractif : il peut créer en quelques lignes de codes des cartes. Comme son nom l'indique il peut accepter simultanément plusieurs fournisseurs de données. Il est possible de rajouter un certains nombre d'options supplémentaires : une barre pour ajuster le zoom, l'affichage des coordonnées, …

En restant toujours fidèle aux logiciels utilisés à l'IFN et ayant fait leurs preuves, il existe une bibliothèque qui surcharge ceux-ci :

- GeoExt[9] : est une librairie javaScript basée sur Ext et sur OpenLayers permettant la mise en place d'applications cartographiques poussées. Le Framework Ext emploie la technologie Ajax (Asynchronous Javascript and XML), ce qui rend les interfaces très réactives, dynamiques et rappelle les interfaces clients riches.

Les principaux outils comme le panel qui gère la construction de la carte (MapPanel), l'affichage des légendes (LegendPanel), l'arbre des layers (treeLayer), …, ont déjà été implémentés, ce qui a rendu la tâche de développement de la maquette beaucoup plus facile et rapide.

[8] Site officiel : http://openlayers.org/
[9] Site officiel : http://www.geoext.org/

Cette librairie n'a pas encore été déployée à l'IFN, seulement les modules originaux Ext et OpenLayers ont déjà fait leur preuve dans des applications officielles. Ce choix a donc été essentiellement motivé par cette connaissance des librairies de base et par de nombreux exemples de scripts existants sur le site GéoTribu[10].

2.3 Représentation des données géographiques

La représentation cartographique constitue un outil privilégié de l'information géographique. Parce qu'elle est immédiate, synthétique, visuelle, elle traduit une information spatiale par une image claire. Nous présenterons dans cette partie deux formes de représentations planes les plus courantes dans les cartes statistiques : les cartes à symboles proportionnels et les cartes choroplèthes. Puis nous présenterons des applications web cartographiques employant diverses technologies et illustrant cette analyse spatiale.

2.3.1 Représentation Plane

La plupart du temps, l'information géographique est réduite à un caractère quantitatif. Ce résumé est réalisé en ramenant la suite ordonnée des valeurs de l'information obtenue à des classes. A l'intérieur de ces groupes la différence d'information n'est pas perceptible. On appelle discrétisation d'une série mesurée ce processus de découpage (passage d'une variable continue en variable discontinue). Cette opération doit permettre de rendre l'information géographique traduite sur la carte avec la meilleure visibilité possible.

Selon les règles de la sémiologie graphique de J. Bertin [Bertin, 1974], la représentation correcte d'une variable résultant d'un comptage doit employer un symbole de taille proportionnelle (comme l'aire d'un disque), tandis que les variables qui ne sont pas additives doivent utiliser une représentation basée sur l'intensité du paramètre (comme la couleur).

- Les cartes à symboles proportionnels (cercles, carrés ou autres figures).

Ces cartes représentent des quantités ou des effectifs : la surface du symbole est proportionnelle a la valeur de la variable quantitative représentée. On dessine par exemple un cercle par modalité dont le centre peut avoir soit un attribut alphanumérique (département, capitale, etc.) soit un attribut géométrique (centroïde).

Si nous voulons implémenter dans une application cette représentation spatiale, il faut toutefois se munir de quelques précautions : par exemple dans le cas où la densité d'une modalité est élevée, la figure peut couvrir presque entièrement la surface de la zone (dans ce cas l'information n'est pas située) ou chevaucher d'autres figures voisines (dans ce cas l'information n'est plus lisible). Le premier problème se rectifie en ajustant le calibrage des tailles des figures : la surface totale de celles-ci ne doit pas excéder 10% des contours géographiques des modalités. Pour le second problème, il existe une technique dite de détourage. Cette méthode consiste à tracer en blanc une circonférence permettant alors de distinguer les petites figures placées sur les grandes. En général cette représentation est réussie visuellement si les contours géographiques sont en petit nombre et répartit dans l'espace de manière homogène.

[10] Adresse du Site : http://ks356007.kimsufi.com/arno/geotribu/node/148

Les cartes à symboles proportionnels peuvent s'enrichir en y ajoutant une seconde variable. La proportion reste l'indicateur de la première mesure quantitative choisie, quant à la deuxième information, de nature quantitative et non additive (comme un pourcentage, une deuxième discrétisation) est représentée soit par un panel de couleur, soit par un graphique (camembert, histogramme). Par exemple la figure 9 illustre l'exploitation des surfaces forestières dans la région Auvergne. Le nord de la région présente une plus grande surface forestière d'exploitabilité facile, contrairement au sud où la part d'exploitabilité très difficile est plus représentée. Sur cette carte, l'échelle des camemberts (taille des cercles) est peut-être un peu imprécise, indiquer la plus petite et la plus grande taille apporterait une lisibilité un peu meilleure.

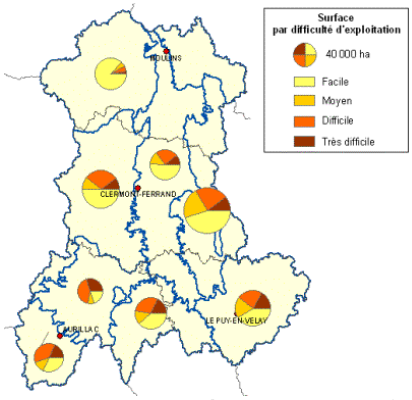


Figure 2-9. Carte des répartitions de la surface par difficulté d'exploitation et par bassin d'approvisionnement.

Une possibilité de confronter deux variables se réalise en accolant deux figures simples (cercles, carrés) pour chacune de couleur différente sur la carte. La corrélation entre elles se lit alors de manière rapide par le rapport de leurs surfaces respectives.

- Les cartes choroplèthes (surfaces colorées ou niveaux de gris).

Elles sont la représentation la plus courante d'une variable numérique (quantité, taux) relative à des zones géographiques. Diverses méthodes de discrétisation existent afin de découper la série statistique en modalités. L'échelle de tons gradués assure alors la visibilité de l'information recherchée.
Cette représentation est le moyen le plus simple à déployer à partir d'un tableau généré par OLAP composée d'une variable de ventilation spatiale et d'une mesure quantitative.

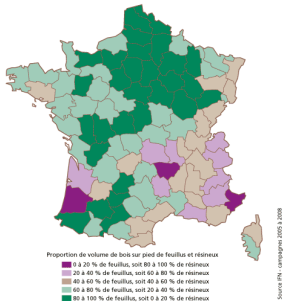


Figure 2-10. Stock de bois sur pied et répartition Feuillus-résineux

Cet exemple illustre le volume de bois sur pied dans les forêts françaises en 2009 (campagnes 2005 à 2008). Elle est extraite du tableau suivant, par une méthode de discrétisation.

REGIONS COTIERES DU NORD-OUEST	242	Peuplement à plusieurs essences feuillues et résineuses	25
REGIONS COTIERES DU NORD-OUEST	453	Indéterminé	11
ILE-DE-FRANCE, CENTRE ET POITOU-CHARENTES	1266	Peuplement à une essence feuillue	109
ILE-DE-FRANCE, CENTRE ET POITOU-CHARENTES	1006	Peuplement à plusieurs essences feuillues	80
…	…	…	…

Tableau 2-4. Volume ventilé par région et composition

Il faut être vigilant sur ce type de carte. Le choix mathématiques de la discrétisation et le choix graphique du panel de couleur ne sont pas anodins : il faut que la combinaison couleur et valeur traduise bien le caractère ordonné. La prise en compte de l'étendue et de l'allure de la série statistique en choisissant un nombre de modalités suffisant afin que les modalités obtenues soient bien réparties dans les principales plages (entre 5 et 10 selon les auteurs) s'obtient par le choix de la méthode de discrétisation. Il en existe plusieurs dizaines classées par branche : les méthodes mathématiques reposant sur les valeurs données, méthodes statistiques ou probabilistes, reposant sur les fréquences, méthodes graphiques, demandant la construction de diagramme ou de courbes auxiliaires [PLUMEJEAUD].

2.3.2 Applications web cartographiques

Il existe un panel d'outils de représentation cartographique mais combien s'appuient sur les entrepôts de données ? La réponse à cette question sera développée dans le paragraphe suivant, mais

nous montrerons ici quelques applications qui proposent une analyse spatiale type décisionnelle avec des technologies différentes.

- Flash, exemple de Géoclip

Flash est un outil de conception d'animations interactives à base de graphismes vectoriels. Ce produit de MacroMédia s'emploie de plus en plus en cartographie. Le format de fichier est vectoriel et propriétaire (les principales spécifications ont cependant été rendues publiques). Afin d'être lu sur internet, l'explorateur doit être équipé d'un plugin d'affichage. Flash permet d'interroger des bases de données, et peut alors afficher des cartes à partir des données géographiques stockées en base.

L'exemple de l'outil Géoclip est intéressant à différents points de vue. C'est un outil web (technologie flash) très interactif qui présente des résultats type décisionnels obtenus par une succincte exploration (plusieurs échelles), exploite de multiples indicateurs (arborescence sur 3 niveaux). C'est outil stocke en base de données l'ensemble des informations nécessaires ainsi que toutes les agrégations possibles suivant toutes les ventilations possibles.

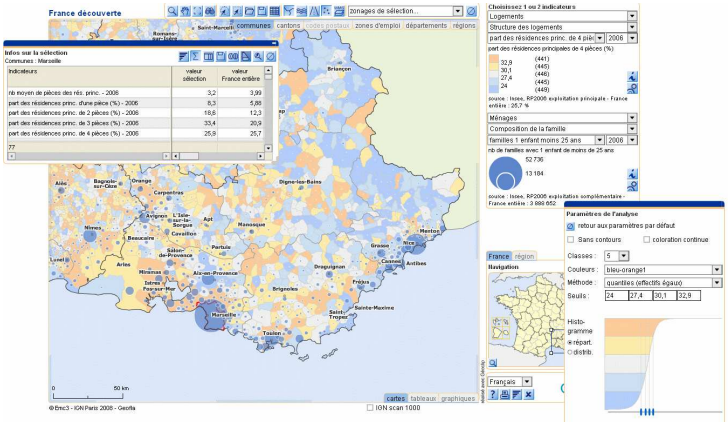


Figure 2-11. France découverte, atlas réalisé par Géoclip de données INSEE

D'autres technologies existent, comme par exemple Java (GeoTools, HyperAtlas), SVG, Ajax (PhilCarto), etc. Nous ne rentrerons pas en détail dans leurs fonctionnalités, cependant elles offrent toutes des possibilités de représentation cartographique interactive.

3 Entrepôt de données géographiques

Nous venons de présenter rapidement deux magnifiques gammes d'outils d'exploitation des données : les applications décisionnelles et les logiciels SIG. Afin de répondre à des questions telles que où se trouvent les plus gros volumes de douglas en France métropolitaines (bois utilisé pour construire son chalet), vous avez aujourd'hui :

- en utilisant les outils SIG à écrire des requêtes complexes, puis avec l'aide d'un expert, à les déployer. Cette opération doit être répétée à chaque nouvelle demande.
- en utilisant les outils BI classiques juste à consulter les tableaux ou rapports, mais ceux-ci sont dépourvus des dimensions et mesures spatiales.

En combinant ces deux gammes de produits naît la technologie du BI Géospatial. L'énorme potentiel de cette alliance permet d'employer simultanément les techniques éprouvées des entrepôts de données, des analyses OLAP, des outils de reporting, des tableaux de bord et du data mining aux techniques d'analyse spatiale et de visualisation des cartes.

Le concept SOLAP (Spatial On-Line Analytical Processing ou processus d'analyse spatio-temporelle interactif), développée par le Centre[11] de Recherche en Géomatique de l'Université Laval (Québec, Canada), s'appuie sur l'ajout de la dimension spatiale au concept de l'OLAP avec adaptation des opérations d'analyses, enrichissement de l'interface et affichage cartographique. SOLAP permet l'interaction dynamique entre des données descriptives et des données géographiques relatives à un territoire.

[11] http://www.spatialbi.com/

Actuellement très peu d'outils couplés de ces deux technologies existent sur le marché (MapInfo (ESRI) avec SAP (Cognos IBM), JMap Spatial avec Kheops[12]), mais aucun n'existe dans le libre (on s'en approche cependant).

Ce dernier paragraphe de cet état de l'art a pour objet de faire découvrir cette nouvelle technologie, et présente quelques développements existants, illustrés par des applications.

3.1 Etat des lieux

Les solutions informatiques du décisionnelles n'ont pas été prévues avec l'optique du spatial : les traitements spatiaux, la navigation avec des données géospatiales ne sont pas supportés. La technologie du business intelligence n'a pas été développée avec la dimension spatiale géométrique. Aujourd'hui, les éditeurs de logiciels décisionnels adaptent l'informatique BI à la création d'un environnement de données propre à la prise de décision sur les objets géographiques.

Habituellement la dimension spatiale est traitée comme les autres dimensions de l'hypercube, c'est-à-dire de façon purement alphanumérique avec l'utilisation des codes ou libellés des noms des lieux hiérarchisés (par exemple le nom du département, de la région administrative, …). Dans ce contexte l'exploration des données spatiales est fortement limitée et l'analyse de leurs interrelations (adjacence, intersection, chevauchement) n'est pas exploitée. La navigation dans l'hypercube reste purement une navigation thématique.

Une alternative afin de rendre plus attractif l'analyse et de voir les relations des phénomènes étudiés sur un territoire, est d'ajouter une propriété géométrique (polygone, centroïde, etc.) à tous les niveaux des hiérarchies de la dimension spatiale. Dans ce cas on dit que la dimension est une dimension spatiale géométrique. La navigation dans le cube devient donc graphique par sa représentation cartographique et tabulaire par sa représentation multidimensionnelle. Cette dimension permet d'observer facilement des phénomènes spatiaux, de les représenter en relation avec d'autres dimensions, de forer interactivement sa granularité.

A ces deux types de dimensions spatiales (données alphanumériques et objets géométriques) s'ajoutent la dimension spatiale mixte. Dans ce cas pour une hiérarchie, les niveaux peuvent avoir soit une propriété textuelle, soit une propriété spatiale. Par exemple, si le premier niveau est un point (codifié par un identifiant numérique), les niveaux supérieurs peuvent être associés à des polygones (contour des régions). L'inverse est envisageable aussi, le premier niveau de la hiérarchie est géométrique et les niveaux supérieurs sont uniquement nominaux.

Les deux dernières dimensions se différencient de la première par la conception de l'entrepôt de données (stockage d'éléments géométriques, manipulation par des fonctions spatiales) mais aussi par son exploitation (représentation graphique). Ce premier point commence aujourd'hui à être traité dans les suites décisionnelles, contrairement au second aspect qui trouve déjà une certaine maturité dans les logiciels du marché.

Dans un modèle multidimensionnel spatial, les mesures peuvent avoir aussi une composante spatiale. Comme nous l'avons vu précédemment dans le tableau 2-1 de ce chapitre, certains opérateurs spatiaux ont des propriétés agrégatives. Mais ces fonctions sont beaucoup plus complexes et différentes que les celles utilisées dans les outils OLAP. Les technologies SOLAP proposent d'utiliser par exemple l'union, l'intersection ou le barycentre. Leurs applications permettent alors d'obtenir des synthèses d'un phénomène spatial.

Par l'association de la représentation cartographique et de la navigation OLAP, l'utilisateur se déplace dans la structure multidimensionnelle et obtient des représentations des données via un affichage cartographique, tabulaire ou en diagramme statistique qui sont des fonctions des dimensions,

[12] http://www.kheops-tech.com/fr/home/index.jsp

des mesures et des niveaux de hiérarchies sélectionnés. Le couplage des fonctionnalités OLAP et SIG a permis l'émergence d'une nouvelle catégorie d'outils d'aide à la décision plus adaptés à l'exploration et à l'analyse spatio-temporelle de ces données. Ces outils ont été regroupés sous la « Technologie SOLAP » [BED 2005].

Les logiciels d'applications cartographiques ont un sérieux inconvénient face à cette nouvelle technologie SOLAP. L'atout de celles-ci provient de la gestion des structures dimensionnelles des cubes de l'environnement décisionnel. En effet une exploitation d'un forage dans un outil SIG n'est en réalité qu'une série de transactions. De plus la constitution des cartes résulte d'un processus dynamique : la carte est associée à une suite successive d'opérations, à l'opposé des logiciels de visualisation de carte, où la carte est associée à une unique analyse. Enfin cette technologie en gérant la représentation graphique, permet à tous les utilisateurs, non experts en cartographie d'afficher les résultats d'analyse sur les cartes.

3.2 Applications OLAP avec SIG intégrés

Les exemples d'applications sont encore assez peu nombreux dans le monde industriel, environnemental, …. Cela s'explique par le fait qu'il y ait assez peu d'offres commerciales, et encore moins d'offres dans le monde des logiciels libres, d'outils intégrés présents sur le marché. Les solutions commerciales déjà évoquées précédemment JMAP-SOLAP (sur la base d'un partenariat technologique entre Kheops) et SAS Web OLAP (sur la base d'un partenariat technologique entre SAS et ESRI) ont été déployées sur le marché en 2005.

Nous vous proposons ici quelques exemples d'applications. Les trois premières sont plutôt classées dans la famille d'outil dont la dominante est OLAP, tandis que les deux dernières applications ont des caractéristiques à dominance mixte (OLAP et SIG).

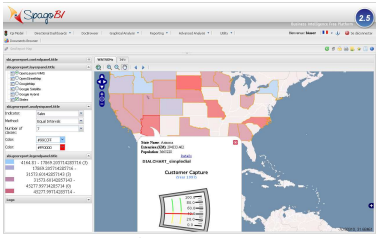


Figure 2-12. Interface Cartographique SpagoBI

- Dans le monde des logiciels libres, l'offre est inexistante. SpagoBI[13], solution de Business Intelligence open Source venu d'Italie, seul projet capable de concurrencer Pentaho, intègre un modèle GIS avec le couple cartoweb/WMS[14].

D'après la démonstration en ligne et sa documentation associée, le stockage des objets géométriques est effectué dans une structure annexe (avec une gestion cependant riche des layers), et le GeoEngine[15] permet aux utilisateurs de manière dynamique de regrouper les informations, selon des hiérarchies géographiques (ex. nation, pays, ville). La barre d'outils offre le choix d'un zoom des clients OGC (sur l'échelle cartographique).

- GeOLAP[16] est un projet réalisé courant 2008 pour une mise en œuvre théorique mi 2009 par Camptocamp utilisant leur framework libre MapFish. Cette application propose une interface interactive cartographique présentant des indicateurs sociodémographiques dont l'utilisateur peut :

[13] http://www.spagoworld.org
[14] Web Map Service
[15] Moteur geospatial de SpagoBI
[16] http://www.geobi.org/2009/01/geolap-geospatial-on-line-analytical.html

- opérer des slices sur la dimension géographique du cube
- réaliser un drill-up sur la dimension cartographique afin de définir la granularité du calcul des indicateurs et de leur unité de représentation
- choisir une seconde dimension pour analyse (carte choroplèthe et à symboles proportionnels)
- paramétrer le choix de représentation.

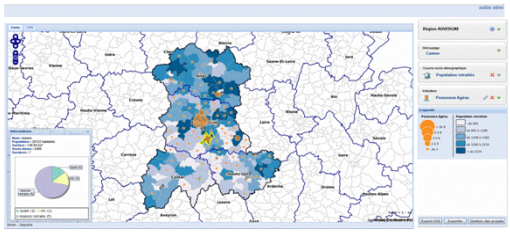


Figure 2-13. Maquette GeOLAP

Attention ce projet a été réalisé sans utiliser les outils de type BI mais il représente un bon cas d'utilisation pour identifier ce qui serait actuellement nécessaire de faire dans les applications avec les technologies décisionnelles. Il manque donc les actions sur la dimension cartographique (recomposition, croisement, subdivision, …).

- DHIIS[17] (The District Health Information System) est un système d'information sur la gestion de la santé en Afrique, open source et très flexible basé sur un entrepôt de données. Il est développé par le programme de santé des systèmes d'information (PSIS) du projet. Ce portail collecte, valide, analyse et présente des données statistiques agrégées, adapté à des activités intégrées de gestion de l'information de santé. Ces capacités SIG sont définies par :
  - Intégration d'un SIG client (MapFish Client) qui permet la cartographie thématique (définir des ensembles de légende, enregistrer des vues personnalisées de cartes préférées).
  - Un gestionnaire des couches cartographiques
  - Génère des cartes via le croisement des informations cartographiques et de l'entrepôt de données.

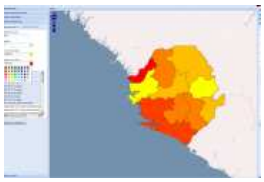


Figure 2-14. DHIS - Cartes thématiques

Figure 2-15. DHIS - Tableau de bord

DHIS est un serveur J2EE (sur une architecture 3-tiers), employant les frameworks Hibernate, Spring, etc. Comme les deux précédents exemples, la source SIG est externe à l'entrepôt de données.

- Un prototype[18] d'une application SOLAP a été réalisé entre le CRC les mines ParisTech et le CRG de l'Université de Laval afin de représenter en ligne la gestion des risques naturels de France métropolitaine. Ce portail s'appuie sur 2 bases de données CORINTE (données des risques naturels) et GEOFLA (base de l'IGN qui contient la géométrie des entités administratives). Deux cubes ont été implémentés « Catastrophes naturelles » et « Plans de préventions ».

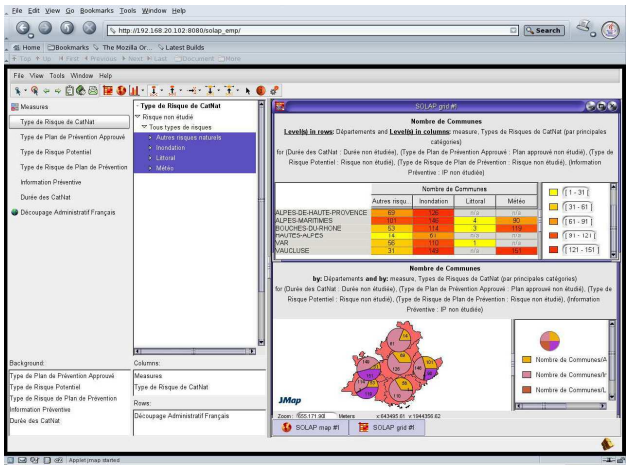


Figure 2-16. Portail Internet sur les risques naturels en France

L'application apporte beaucoup de fonctionnalités supplémentaires aux analyses traditionnelles par la manipulation et la visualisation des données spatiales. Les volets de résultats

[17] http://www.dhis2.com/
[18] http://www.crc.mines-paristech.fr/fr/solap.html

tabulaire, cartographique et graphique apparaissent simultanément suite aux choix d'exploration dans le cube, des outils permettent de naviguer entre les différents niveaux des dimensions.

- GeWOlap[19].

SOLAP comporte cependant quelques inconvénients. Par exemple un dès inconvénient est que la dimension spatiale est décrite par des hiérarchies dont les membres sont des objets géographiques liés par des relations topologiques d'inclusion ou d'intersection. Cette définition ne reflète pas la sémantique sous-jacente aux liens hiérarchiques. En effet, les objets géographiques peuvent être en relation avec d'autres objets à travers des relations spatiales, des relations de généralisation et des relations non spatiales. La prise en compte de ces types de relations peut-être fondamental dans l'analyse multidimensionnelle, car à chaque type de hiérarchie correspond une analyse différente, qui peut se traduire en différentes politiques d'agrégation et de navigation. [BIMONTE]

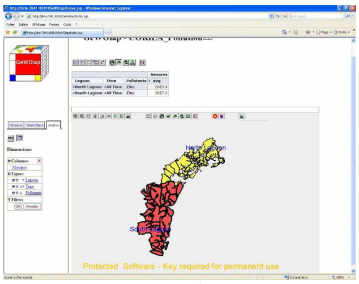


Figure 2-17. Interface GeWOlap

En 2007, l'équipe de LIRIS[20] définit le concept d'OLAP Géographique [TCH et BIM]. Cette approche reformule les concepts du Spatial OLAP afin de prendre en compte la composante spatiale et sémantique de l'information géographique et la flexibilité de l'analyse spatiale.
Cette étude a permis de concevoir un prototype GeWOlap qui implémente les opérateurs multidimensionnels définis par GeoCube (modèle formel et algèbre associée).

Cette solution web couplant le OLAP et SIG gère les mesures et dimensions géographiques complexes. Elle permet une analyse spatio-multidimensionnelle à l'aide d'une interface tabulaire synchrone à la carte interactive couplée à des affichages graphiques.

3.3 À venir GeoMondrian et SOLAPLayers

Spatialytics.org[21] est un site où la communauté peut bénéficier et participer aux récentes initiatives Open Source en GéoBI, autour des trois principaux projets que sont GeoKettle, GeoMondrian et SOLAPLayers. Le site vise à mettre en valeur son offre technologique en BI et Géospatial Open Source auprès et avec des entreprises compatibles technologiquement notamment celles qui exploitent Kettle et/ou Mondrian, PostGIS et/ou GeoExt, …. Les projets disponibles sont des logiciels gratuits et dont le code source est libre.

Nous nous intéresserons aux composants de GéoBI : GeoMondrian et SOLAPLayers. Ils sont disponibles en version de développement (les projets sont toujours en démarrage, donc n'ont pas encore de version stable) et sont téléchargeables sur le site :

- https://geomondrian.svn.sourceforge.net/svnroot/geomondrian
- https://spatialytics.svn.sourceforge.net/svnroot/spatialytics

En mars 2010, Spatialytics confirmerait un partenariat technologique avec SpagoBI. Les possiblités de collaboration sont nombreuses. SpagoBI utilise déjà le serveur Mondrian et est sensible à la liaison entre le BI avec la cartographie (voir §3.2 de ce chapitre).

- GéoMondrian.

GeoMondrian est le premier serveur open source spatial Olap. Basé sur Mondrian, le moteur OLAP OpenSource intégré dans Pentaho, geoMondrian permettra de stocker dans un cube des données spatiales (une cartouche spatiale olap) et non plus dans une structure annexe (base de données,

[19] http://eric.univ-lyon2.fr/~sbimonte/doc/TheseSandroBimonte.pdf
[20] LIRIS : Laboratoire d'InfoRmatique en Image et Systèmes d'information
[21] http://www.spatialytics.org/

services web ou fichiers shapes externes) comme c'est le cas bien souvent. Aujourd'hui, GéoMondrian ne supporte que le SGBD PostGIS pour l'entrepôt de données.

GeoMondrian est une réalisation de l'équipe de recherche du groupe GeoSOA à l'université de Laval au Quebec. Les termes de la licence de GeoMondrian correspondent à ceux de la licence EPL (Eclipse Public License).

Il implémente les types de données géométriques, ce qui permet le stockage des propriétés géométriques des membres des hiérarchies des dimensions spatiales et ainsi de filtrer les membres suivant les prédicats topologiques (intersect, contains, within, …). Il ajoute aussi des mesures dans le cube de données dont les calculs sont basés sur des attributs scalaires provenant de caractéristiques spatiales (surface, distance, …). Afin d'exploiter ces nouveaux objets, des extensions au langage d'interrogation MDX ont été ajoutées. L'exploration multidimensionnelle peut alors profiter de ces nouvelles capacités.

Ci-dessous un exemple de requête utilisant la fonction distance comme filtre spatial sur les membres d'une dimension.

```
Select {[Measures].[Population]} on columns,
          Filter (
       {[Unite geographique].[Region economique].members},
              ST_Distance([Unite geographique].CurrentMember.Properties("geom"),
              [Unite geographique].[Province].[Ontario].Properties("geom")) < 2.0
       ) on rows
From [Recensements]
Where [Temps].[Recensement 2001 (2001-2003)].[2001]
```

- SOLAPLayers.

SOLAPLayers est un serveur web open source, capable d'afficher des cartes dont les résultats sont directement issus d'une analyse OLAP de GéoMondrian.

Il permet surtout une navigation géospatiale dans le cube de données à l'aide du serveur GéoMondrian. Ce composant cartographique vise aussi à être intégré dans différents frameworks de tableau de bord afin de produire de véritables tableaux de bord géo-analytiques interractifs. Il permet encore la représentation cartographique de mesures et de membres d'une dimension spatiale sous la forme de cartes choroplèthes (intervalles fixes ou intervalles égaux dynamiques).

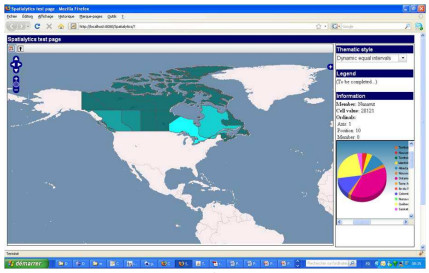


Figure 2-18. Interface de Spatialytics

SOLAPLayers utilise les librairies OpenLayers et Dojo et peut, évidemment se connecter au serveur spatial GeoMondrian.

GeoMondrian et Spatialytics ont déjà été adoptés par certaines organisations, comme par exemple le projet GeoBI (http://www.geobi.org). GeOLAP n'est autre que GeoMondrian et GeoReport devrait bénéficier bientôt des capacités de Spatialytics (un étudiant financé par le programme Google Summer of Code 2009, sous l'égide de l'OSGeo, travaille actuellement à cette tache à l'Université Laval) [BALIZ-MEDIA.com]

3.4 Synthèse

Ce chapitre a permis de rassembler les éléments permettant de définir le géodécisionnel. Comme nous avons pu le constater c'est un domaine encore émergent qui n'a pas encore fait l'objet

d'une véritable appropriation par le monde économique des entreprises comme le secteur de l'environnement par exemple.

Les systèmes d'information géodécisionnels reposent avant tout sur les fondements des systèmes d'information décisionnels. Ceux-ci sont utilisés comme une surcouche des systèmes d'information opérationnels fonctionnant sur un mode transactionnel.

La BI a fait ses preuves en termes d'efficacité et de performance pour l'analyse à partir des sources de données hétérogènes. L'handicap est que ces systèmes décisionnels ne prennent pas en compte la composante géographique des données. Les technologies de la géomatique au travers des SIG ont atteint eux aussi un degré de maturité dans les entreprises. Mais la géomatique repose sur des processus OLTP n'intégrant pas des fonctionnalités permettant d'obtenir des vues de synthèse comme le permettent les processus OLAP.

Le géodécisionnel est la fusion entre le décisionnel et la cartographie web. Il s'agit de l'enrichissement de la composante spatiale dans la modélisation multidimensionnelle avec les dimensions spatiales et les mesures spatiales. Des exemples d'applications et de logiciels OLAP-SIG intégrés ont été présentés notamment à travers la technologie SOLAP. Dans l'open source nous avons vu que les choses bougeaient également avec l'aide d'une communauté dynamique : Spatialytics.

Un juste milieu afin de réaliser une étude sur les entrepôts de données spatiaux à l'IFN est à trouver entre les solutions à caractères OLAP dominant et les travaux de recherche sur des analyses complexes spatiaux-multidimensionnels. L'étude peut dans un premier temps se pencher sur la gestion des objets géométriques en tant que tel afin de permettre de réaliser des forages spatiaux, des pivots et des coupes géométriques de manière interactive sur les cartes.

Après avoir rassemblés tous les éléments permettant d'expliciter ce que l'on entend et ce qui existe dans le monde du géodécisionnel, il convient maintenant d'étendre le travail de modélisation à un cas pratique et dans la conception d'un entrepôt de données géospatial traitant des besoins d'exploration et dans une solution logicielle proposant aux utilisateurs un certain nombre de fonctionnalités SOLAP. C'est l'objet des chapitres suivants.

Analyse de l'existant et proposition

Dans ce chapitre, nous rapporterons les motivations qui ont conduit à concevoir une étude de faisabilité visant à mettre en place un entrepôt de données avec une partie des données de l'inventaire forestier. Afin de comprendre le propos de notre étude, nous décrirons une des missions de base de l'IFN qui consiste au suivi statistique de la forêt française et pour cela nous présenterons sa méthode de calcul des résultats produits. Enfin, afin de mieux se rendre compte des besoins actuels, nous brosserons une courte analyse technique de l'architecture du système d'information, plus particulièrement la chaîne de traitement qui *in fine* produit des résultats statistiques.

1 Présentation des motivations

Nous présenterons ici les désirs exprimés par les ingénieurs forestiers de l'IFN en ce qui concerne les besoins pratiques, les contraintes ou les idées techniques concernant l'analyse et la valorisation des données de l'inventaire. De certaines de ces envies vont naître la mise en place de l'entrepôt de données et la conception d'une maquette d'une application type SOLAP que nous présenterons largement dans les chapitres 4 et 5.

1.1 Pour une mise en place d'un entrepôt de données

Dans l'ensemble du système d'information de l'Inventaire forestier, la chaîne de traitement de l'information permet de mettre à disposition rapidemment et sous une forme conviviale les résultats produits et validés. Basée sur une architecture orientée service (SOA) (nous la détaillerons dans le troisième paragraphe), les principaux modules sont la base d'exploitation contenant les données exploitables, le service de calcul, qui permet d'établir à la volée l'estimation d'un résultat avec un intervalle de confiance et l'application OCRE[22], interface web qui permet de produire par une démarche séquentielle en huit étapes des résultats statistiques sous formes de tableaux à deux dimensions et extractables sous un tableur.

Malgré cette facilité informatique, l'utilisateur doit savoir ce qu'il cherche et maîtriser les données et les requêtes qu'il utilise. L'approche choisie lors de la refonte de l'ensemble du système d'information en 2005 se décline de l'interface du site internet, dédiée à tous les internautes demandeurs de résultats forestiers : c'est une succession itérative d'étapes dont l'ordre est important (choix du domaine spatial, choix du domaine temporel, choix des objets d'inventaire, restriction du domaine d'étude, choix des variables, choix des critères de ventilations). Ce cheminement n'est pas approprié pour la recherche exploratoire d'une information qui consiste à effectuer différentes analyses, la restitution des résultats amenant à refaire une nouvelle recherche.

Le service de calcul s'appuie sur une base de données relationnelles dont le modèle a été ajusté afin d'optimiser ses performances (duplication de données, introduction de redondances, index, …). Les résultats, calculs de variables et variances, vus la complexité des calculs, restent performants après cinq campagnes d'inventaire (de l'ordre de quelques secondes).

Le service de production des résultats de l'inventaire suit une recherche multidimensionnelle, de façon similaire aux opérations d'analyse en-ligne OLAP. Ces deux domaines d'applications s'intéressent à la production de résumés statistiques moyennant les dimensions d'ensemble de données. La structure conceptuelle d'un entrepôt de données et le traitement des données statistiques forestières ont exactement les mêmes composants : une mesure, une fonction, une ou plusieurs dimensions, zéro ou plusieurs hiérarchies.

[22] OCRE : Outil de Calcul de REsultats. Application interne de l'IFN.

Les bases de données statistiques touchent le monde des statisticiens, et concernent plutôt les domaines d'applications comme le recensement, l'analyse de données économiques, études des ressources naturelles, …. Le domaine des applications OLAP et des entrepôts de données ont été mis en place principalement dans le monde de la finance, du commerce pour répondre à des demandes d'analyses ou à des demandes de tableaux de bord, sur de gros volumes de données. Classiquement utilisés pour offrir une aide à la décision, ils permettent de déterminer des indicateurs de suivi et des analyses en temps réel.

Dans le cadre des activités de l'IFN, en réussissant à modéliser l'ensemble des processus : la stratification (opération de découpage du territoire en strate[23]), la gestion des campagnes d'inventaire et les dimensions géographiques, les entrepôts de données par leur mise en place pourraient servir de base :

- à la production de résultats
- à apporter des interfaces orientés navigation : tableaux de bord, cartes et graphes.
- à offrir aux utilisateurs une exploration très interactive dans des tableaux multidimensionnels
- etc.

OLAP est bien adapté aux calculs complexes, aux cumuls selon des structures organisationnelles compliquées et aux calculs qui portent sur des lignes différentes. La technologie semble assez puissante pour y intégrer les estimateurs de l'IFN et ce nouvel entreposage des données accessible à tous faciliterait l'utilisation des données.

Les technologies du monde décisionnel sont proches de celles utilisées à l'IFN. Elles comportent des bases de données relationnelles, des architectures client-serveur, la modélisation des métadonnées, des interfaces graphiques, …. Les utilisateurs, aussi bien les experts que les amateurs, s'adapteraient facilement à cette nouvelle approche, et les coûts d'investissements, grâce aux outils existants dans l'open source, n'augmenteraient pas de beaucoup.

1.2 Vers une mise en place d'une application type SOLAP

Certains utilisateurs internes expriment depuis longtemps leur désir d'avoir une application qui permettrait d'analyser et d'explorer les données de l'inventaire par le biais d'une **application interractive, conviviale, rapide et visuelle**. En effet, les données dendrométriques, écologiques et cartographiques de l'inventaire obéissent à une recherche multidimensionnelle et donc l'organisation de leur structure en entrepôt de données spatial permettrait une analyse multidimensionnelle consistant par exemple à explorer ces données d'un niveau détaillé à un niveau agrégé, ou l'inverse, grâce aux outils OLAP.

L'étude sur les données via une interface type JPivot devient très facile et rapide. Par exemple, afin de publier des résultats cohérents par rapport à leur intervalle de confiance, il est intéressant de pouvoir zoomer jusqu'à un niveau de détail intéressant pour les analystes et si le niveau détaillé est trop faible, pouvoir ré-agrégé. Une autre situation se présente quand on veut comparer les données dans le temps ou dans leur mode de calcul (choix d'une stratifaction différente, choix d'une autre campagne, modifier son choix de l'estimateur sur l'ensemble des tranches temporelles, …). Comme exemple supplémentaire, les études de ressources doivent concevoir une segmentation de la ressource forestière sur un territoire donné. Cette opération est importante et le choix des meilleurs critères de ventilations est déterminant pour de bons résultats, il faut donc un outil qui permette d'évaluer rapidemment les différentes clés de répartition et leurs validités.

Un besoin des utilisateurs serait d'avoir un outil de visualisation spatiale des résultats produits. Aujourd'hui seule une visualisation des données brutes[24] ponctuelles existe sous forme de cartes et de tableaux. Un requêteur permet de filtrer les données recherchées et les points de l'échantillonage

[23] Une **strate** est un ensemble de points regroupés ayant certaines propriétés en commun.
[24] « Visualisez les données brutes de l'IFN »

s'affichent alors sur la carte, les critères de ventilations s'affichent quant à eux dans un tableau. Générer des cartes avec des aplats de couleurs pour des entités géographiques sans demander à des cartographes apporteraient une indépendance aux utilisateurs. De plus l'éventail des cartes à produire pour apprécier la spatialisation du calcul est large, ce qui rend la tâche longue, complexe et routinière. Et si de surccroît l'intérêt est de décliner la cartographie selon un nouveau critère toute la démarche est à refaire.

L'IFN est fortement engagé sur la thématique des indicateurs de gestion durable, tant au niveau national, qu'européen. Dans cette exploration d'applications décisionnelles, un type d'application type **tableau de bord** spatial impliquant des calculs d'indicateurs spatiaux au plus près des bases de données répondraient à des besoins utilisateurs.

Afin de répondre à ces besoins d'améliorations ou nouveaux des logiciels existants, une maquette d'une application s'appuyant sur un entrepôt des données géographiques et non géographiques, offrant des interfaces d'exploration interractives et intuitives et enrichies d'une visualisation des données agrégées sous forme de carte a été réalisée et a permis de confronter quelques utilisateurs internes de l'IFN à ce nouveau type de solutions.

2 La méthode d'inventaire

Dans cette section, nous décrivons la méthode statistique de l'inventaire. Afin d'intégrer au mieux les données dendrométriques dans un entrepôt, il est important de bien la comprendre. Beaucoup d'éléments essentiels y sont introduits et constituent les fondements de la modélisation multidimemsionnelle (système d'échantillonnage, stratification, objets inventoriés, poids des points, …).

2.1 Sondage statistique

L'Inventaire Forestier National (IFN) est chargé de réaliser l'inventaire permanent du patrimoine forestier sur tout le territoire métropolitain. Il acquiert au fil du temps des données pour toutes les forêts publiques ou privées, il estime la ressource forestière française, détermine son évolution, calcule des indicateurs pour la gestion durable, établit la carte forestière, etc.

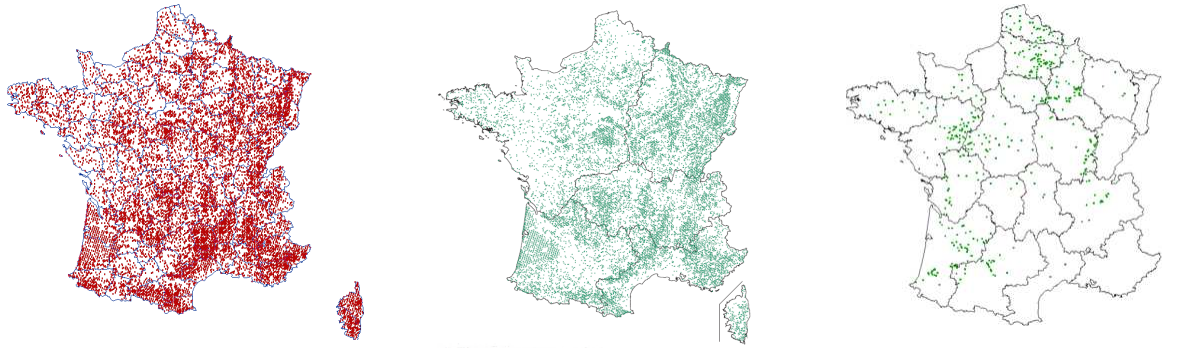
Depuis 2005, un an après l'INSEE[25], la méthode de l'inventaire a changé. Au lieu de visiter chaque département tous les douze ans, l'IFN réalise chaque année un sondage spatial systématique. L'échantillon est constitué pour une période de dix ans sur l'ensemble du territoire français. Un dixième des opérations d'inventaire sont réactualisées chaque année sur tout le territoire.

Sur cet échantillon de base, on le subdivise afin de constituer des sous-échantillons adaptés aux différentes opérations réalisées (photo-interprétation, reconnaissance et levé) et aux différentes formations inventoriées (forêt, peupleraie, lande, …).

La première phase de travail est la photo-interprétation des placettes centrées sur les points d'inventaire qui définit l'échantillon dit de phase 1. Cet échantillon complet comporte environ 80 000 points d'inventaire par an. On y détermine des données telles que la couverture du sol, son utilisation, l'accessibilité, ….

A partir de cet échantillon, on détermine le sous-échantillon des points à visiter. Une fraction annuelle des points levés totalise environ 8000 points. Les points sont repérés par leurs coordonnées géographiques. Les agents y relèvent des observations concernant le peuplement forestier, des informations sur la végétation et aussi des mesures sur les arbres (hauteur, diamètre, …). L'échantillon des arbres est lié à l'échantillon des points : par an on relève un effectif de 65000 arbres vifs. Le recueil des saisies et des observations totalisent par point environ 400 variables.

[25] Institut national de la statistique et des études économiques



Echantillon 2005 | Les points levés en forêt de production en 2005 et 2006 | 500 points peupleraies 2005, 2006 et 2007

Figure 3-1. Quelques échantillons de points d'inventaire

Cette méthode permet de produire des résultats au niveau national tous les ans. Cependant, afin de fournir des résultats statistiques fiables au niveau d'une région administrative, les estimateurs doivent cumuler les échantillons sur cinq campagnes d'inventaire.

Les campagnes d'inventaire sont annuelles et couvrent uniformément le territoire. En ajoutant les échantillons de différentes fractions, on augmente le recueil d'informations, et par conséquent on améliore les résultats produits. Les résultats pluriannuels sont actuellement déduits par une moyenne simple des estimateurs pour chacune des subdivisions temporelles.

Ainsi la compilation des observations de deux campagnes annuelles France entière (2005 et 2006) a permis de fournir des résultats nationnaux et interrégionnaux avec une précision statistique supérieure à ceux fournis en 2005 à l'issue d'une seule campagne.

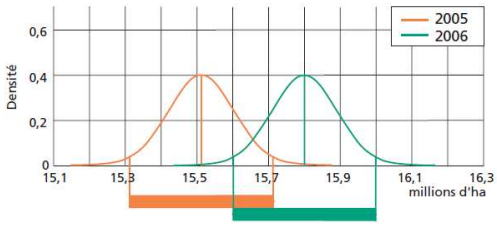


Figure 3-2. La superficie forestière en 2005 et en 2006(5)

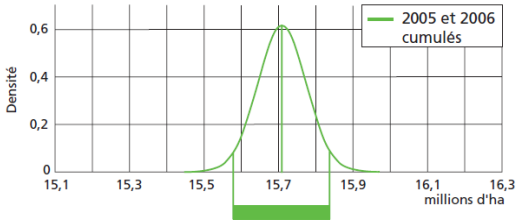


Figure 3-3. La superficie forestière à partir des mesures de 2005 et 2006([26])

Par exemple en 2005 la superficie forestière était estimée à 15,5 millions d'hectares avec un intervalle de confiance de plus ou moins 0,2 million d'hectares au seuil de 95%. Cela signifie qu'il y a une probabilité de 95% pour que la superficie forestière en France en 2005 soit comprise entre 15,3 et 15,7 millions d'hectares. En 2006, la superficie forestière est estimée à 15,8 d'hectares, avec ce même intervalle de confiance de plus ou moins 0,2 million d'hectares au seuil de 95%. Les courbes de la figure 3-2 représentent l'incertitude sur la superficie de chaque échantillon annuel.

En cumulant les deux échantillons annuels, on augmente la taille de l'échantillon, ce qui améliore la précision des estimations, en réduisant l'intervalle de confiance (figure 3-3). La superficie forestière est alors estimée à 15,71 millions d'hectares à plus ou moins 0,13 million d'hectares au seuil de 95%. Autrement dit la superficie forestière peut valablement être supposée comprise entre 15,58 et 15,84 millions d'hectares. [IFN 2006]

[26] Extrait « La forêt française – Les résultats issus des campagnes d'inventaire 2005 et 2006 », IFN-2007.

Les dates ont des rôles différents dans une publication de résultats et sont toutes importantes. On parle :

- Campagnes d'inventaire : ce sont les années de campagnes qui vont constituer l'échantillon de points.
- Année millésime M : c'est l'année de référence statistique. Elle peut être représentée par exemple par la moyenne des années de campagnes.
- Date de validation du résultat : cette date est postérieure aux deux précédentes. Elle indique quand le feu vert est donné afin de publier les résultats. Les résultats ne sont plus modifiés après cette date.

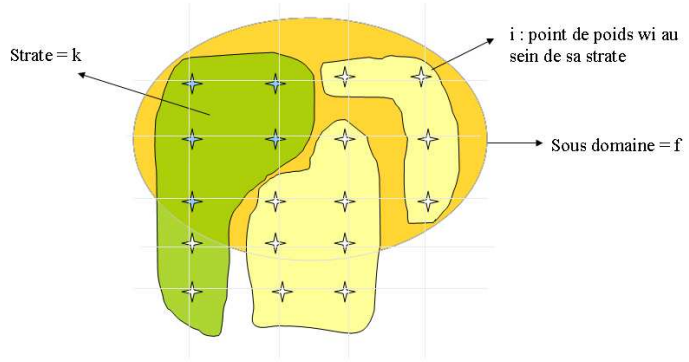
Un énoncé peut s'avérer très vite long à écrire. Par exemple le volume total de bois pour la France millésimé 2007 entrant en vigueur le 15 septembre 2010, calculé avec les campagnes d'inventaire de 2005 à 2009 (moyenne des 5 résultats annuels) est estimé à 2 423 millions de mètre cubes (Mm³) avec un intervalle de confiance à 95% de 40 Mm³ soit 1,6%.

2.2 Estimation des résultats

Les calculs des estimateurs à l'IFN sont des calculs statistiques. La méthode de calcul repose sur une post-stratification. Cette étape intervient après la collecte des informations. La stratification améliore l'efficacité du sondage et donc la précision des résultats.

Le territoire métropolitain est découpé en strates par des données cartographiques (carte des départements, carte des propriétés, carte forestière simplifiée) puis par des données de phase 1 On connait ensuite la surface de ses strates, elles ne sont jamais vides, et chaque point d'inventaire est affecté à une strate donnée.

La production des résultats statistiques reposent sur le choix de cette stratification. La figure ci-dessous résume ce partionnement :



Une strate est un ensemble de points regroupés ayant certaines propriétés en commun (par exemple on divise le territoire en strates ayant le même département, type de formation végétale (FAO) et type de propriété). Les strates permettent de pré-calculer des poids statistiques pour des ensembles de points (c'est la stratification).

Figure 3-4. Schéma simplifié d'une stratification

Les types de résultats que nous cherchons à calculer, concernent les estimations de surface, par comptage de points, des totaux, par somme de variables quantitatives sur les points et des moyennes, rapport des deux grandeurs précédentes (on parle de moyenne spatiale, par exemple le volume à l'hectare).

Ces résultats sont déterminés sur un sous-domaine. On appelle sous-domaine f une extraction du territoire restreint à certains critères qualitatifs relatif à la placette, par exemple combinaison de critères cartographiques et de critères de reconnaissance (phase 1).

Certains résultats peuvent être ventilés par des données qualitatives de type « arbre », comme par exemple le volume sur pied par les diamètres des arbres, leurs essences, leurs formes du houppier, …. Tous les résultats ne peuvent pas pour autant être ventilées par ces données (un arbre n'a pas du surface, un arbre n'a pas de hauteur dominante, …). La notion d'arbre appartient à un point et non pas

à son emplacement géographique. C'est impossible de croiser les estimateurs quantitatifs « peuplement » par des variables qualitatives « arbre ».

La méthode générale de l'IFN adopte des variantes suivant le type d'inventaire dont elle s'occupe. Il s'occupe d'inventorier les forêts de production, les peupleraies (leur protocole différent en 2005, à quelques informations près est aujourd'hui identique à celui des formations boisées), les bosquets, les landes et les ligneux hors forêt. Ces derniers n'ont pas été pris en compte dans l'entrepôt de données. Tous ces objets d'inventaire comportent des protocoles différents, les données relevées leur sont propres. Cependant la méthode de calcul des estimations de chacun reste identique à l'exception des ligneux hors forêt.

En résumé, nous venons de voir trois notions qui impactent le calcul des estimateurs : poids des points dans l'échantillonnage, la stratification, et le sous-domaine d'étude.

Soit f un sous-domaine d'étude, $K = \bigcup_{k \in K} k$ l'ensemble des strates du territoire, S_f la surface du sous-domaine d'étude recherchée, w_i le poids du point i de l'échantillon, on a alors :

$$S_f = \sum_{k \in K} \left(S_k * \frac{\sum_{i \in f \cap k} w_i}{\sum_{j \in k} w_j} \right)$$

Équation 3-1. Surface d'un domaine d'étude

$$T_f(y) = \sum_{k \in K} \left(S_k * \frac{\sum_{i \in f \cap k} w_i}{\sum_{j \in k} w_j} * \frac{\sum_{i \in f \cap k} w_i * y_i}{\sum_{j \in k} w_j} \right)$$

Équation 3-2. Total de la variable quantitative y d'un domaine d'étude

Les mesures de l'IFN, ainsi que leur coefficient de variation[27], sont des mesures dites exhaustives, cela veut dire qu'en stockant comme données : le poids d'un point et le carré du poids, dans l'entrepôt de données on ne perd pas d'information. Théoriquement cela garantit l'agrégation des estimateurs et des intervalles de confiance.

3 Dix giga-octets de données en ligne[28]

Nous décrirons ici l'architecture du système d'informations de l'IFN, en partie les bases de données et surtout la base des métadonnées, base sur laquelle la constitution de l'entrepôt s'est appuyée. Nous présenterons aussi rapidemment les applications en fin de chaîne de traitement qui permettent aujourd'hui de produire des résultats, afin d'analyser les atouts éventuels par l'approche décisionnelle.

3.1 Architecture du système d'information

L'architecture du système d'information de l'IFN repose sur quelques grands principes : la production est séparée de l'exploitation (bases, réseau) pour des raisons de sécurité, de rigueur et de performances, l'architecture est orientée services (l'utilisation de la base d'exploitation peut se faire sans avoir la connaissance de son schéma, les applications métiers sont mises en services, comme le calcul, la stratification), des espaces dédiés aux utilisateurs ont été mis en place afin de leur donner une plus grande souplesse et une autonomie.

[27] Le coefficient de variation d'une variable est le quotient de l'écart-type par la moyenne exprimé en pourcentage.

[28] Titre d'un article écrit par Jean Wolsack, directeur technique de l'Inventaire Forestier jusqu'en 2005.

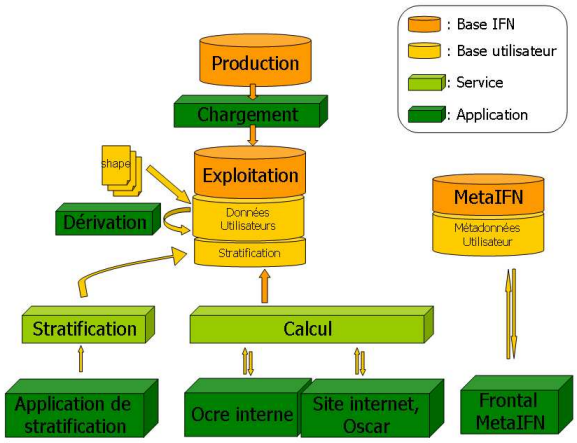


Figure 3-5. Architecture de la chaîne de traitement

Au sein de son système d'information, l'IFN dispose en production de quatre bases de données contenant : les métadonnées, l'historique des opérations d'inventaire, les données collectées sur le terrain et la carte forestière.

L'ensemble de ces données sont traitées par extraction et chargée dans une base d'exploitation annuellement. Cette base a pour objectif de répondre aux besoins des utilisateurs de résultats. Elle contient toutes les données calculées, des données propriétaires aux utilisateurs, des données exogènes, ….

Accessible depuis l'extérieur, sa structure évolue selon les besoins, et son contenu est voué à augmenter régulièrement.

Un service de stratification (permet de réaliser les opérations de stratification suivant le choix des données cartographiques) et un service de calcul (permet de calculer des estimateurs) permettent de fournir des résultats statistiques complets et validés (dans le traitement et dans la mise en vigueur). Ces deux services sont écrits en Java et offrent une efficacité dans l'intéractivité demandée par les utilisateurs.

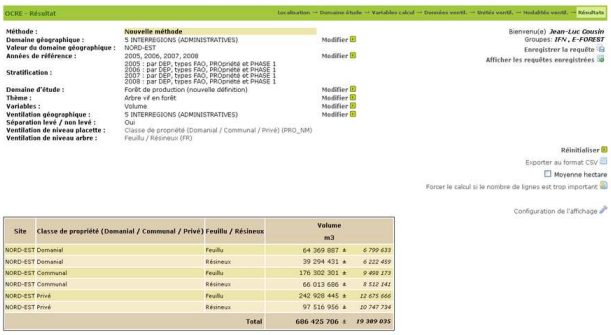


Figure 3-6. Interface OCRE : tableau final

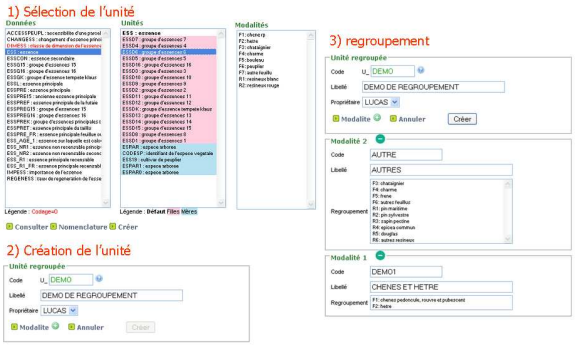


Figure 3-7. Interface Regroupement d'Unité

Des applications web, écrites en Ajax[29], permettent aux utilisateurs de manipuler et de sortir des résultats. Par exemple, l'application OCRE (figure 22) dont nous avons déjà parlé au début de ce chapitre permet aux utilisateurs d'obtenir des résultats de manière itérative.

Une seconde application, « Application de Regroupements » permet de gérer l'ensemble des données dites regroupées. De nombreuses données nominales (dont la caractéristique est que la liste de ses valeurs est finie) ou discrètes (les classes de valeur sont issues d'une donnée continue discrétisée) ont une liste de valeurs possibles trop grande pour une publication de résultats. Il existe une possibilité de regrouper ces ensembles de valeurs en de plus grands sous-ensembles.

Par exemple les classes de diamètre issues d'une discrétisation prennent des plages de valeurs d'étendue de 5 cm de 7,5cm à 250cm et plus, ce qui en donne une cinquantaine. A ce niveau les résultats sont trop détaillés. L'application de regroupement permet de créer alors 4 classes : la classe « Petit bois » regroupera les classes de 2,5 cm à 20 cm, la classe « Bois moyen » regroupera les classes de 22,5 cm à 45 cm, etc …. Derrière cette notion de regroupement, on devine la notion de hiérarchie des dimensions des cubes OLAP.

[29] Ajax : Asynchronous JavaScript And XML

3.2 Les métadonnées à l'IFN

Parmi les bases de données de l'Inventaire Forestier, la base des métadonnées permet de décrire toutes les données (saisies, cartographiques, calculées, applicatives, systèmes, …) sur lesquelles le système d'information s'appuie pour fonctionner. C'est la base de données de référence des données techniques de l'IFN.

C'est un système formel qui fournit l'information d'autorité sur la sémantique (le sens) et la structure de chaque donnée technique. Pour chaque donnée, MetaIFN en donne la définition (selon la nomenclature en vigueur), les qualificatifs qui lui sont associés, l'utilisation qui en est faite dans le système d'information.

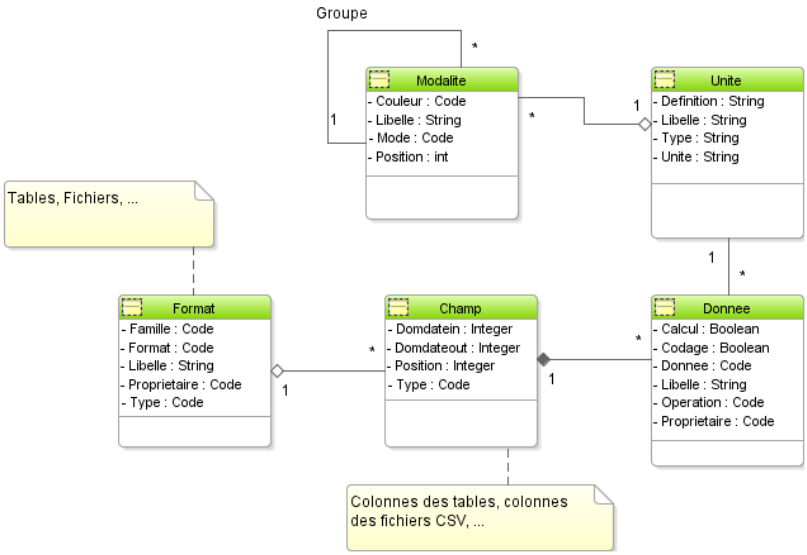


Figure 3-8. Extrait du diagramme de classes des métadonnées de l'IFN

Un extrait du diagramme de classe des métadonnées est illustré dans la figure 22. La classe principale est la donnée. Elle est qualifiée par une description (libellé, définition), par un type (qualitative, quantitative, …), par un format de stockage. Chaque donnée a une unité de mesure. Pour les données qualitatives, l'unité est composée d'une liste de modalités référencées. Celle-ci peuvent être groupées afin de définir des niveaux. Cependant la notion de hiérarchie n'existe pas et rejoint un besoin connu de la direction technique.

Les métadonnées des évolutions des protocoles, des données, des modalités au cours du temps y sont aussi référencées.

Elle constitue le socle de toutes les applications de la chaîne de saisie et de la chaîne de traitement de l'IFN. La base des métadonnées a été d'un grand soutien dans la construction de l'entrepôt de données et des cubes.

3.3 Conclusion

Ce chapitre constitue un préambule aux deux chapitres suivants : il a permis de rassembler les éléments permettant de définir une partie de l'architecture du système d'information de l'Inventaire Forestier National. Les outils développés et maîtrisés permettent aujourd'hui de garantir une qualité et une cohérence des résultats produits et de les procurer aux utilisateurs adaptés à leurs besoins.

Il a montré au début de ce chapitre en quoi les entrepôts de données apporteraient aux utilisateurs et pourraient trouver sa place dans ce processus : production de résultat, accès à des interfaces orientés navigation, offrir aux utilisateurs des représentations géodécisionnelles.

La méthode d'inventaire et le processus de traitement de la chaîne d'information conduisent à fournir des résultats statistiques sur la forêt française, une des missions de l'IFN. La présentation du calcul des estimations des résultats a permis de mettre l'accent sur sa complexité et surtout sur l'aspect « temps » qui n'est pas géré classiquement (les résultats annuels se calculent sur un échantillon pluriannuel se renouvelant continûment par fractions annuelles).

Après avoir expliciter ces outils de travail, il convient maintenant de mettre en place l'entrepôt de données, le modéliser, l'implémenter puis d'intégrer un serveur OLAP pour enfin étudier plus précisément les atouts du décisonnel voire du géodécisonnel. C'est l'objet des deux prochains chapitres.

Entrepôt de données à l'Inventaire Forestier.

Ce chapitre a pour objectif de présenter la démarche mise en œuvre pour modéliser les données de l'Inventaire Forestier National et mener à bien l'utilisation d'outils d'exploration. L'objectif premier est d'aboutir à la construction de l'entrepôt de données, pour ensuite démontrer le potentiel méthodologique et technologique des systèmes d'information géodécisionnels. Etant donné qu'il s'agit d'une étude de faisabilité et donc d'une démarche expérimentale, le travail commence par une étape de modélisation, puis d'une première implémentation qui pourra dans une seconde version s'améliorer et s'optimiser, ce qui sera détaillé en dernier lieu.

1 Modélisation dimensionnelle des données de l'Inventaire Forestier

Au chapitre 2, les concepts de dimensions hiérarchiques, de mesures et d'opérations ont été présentés. Ce paragraphe décrira comment dans l'entrepôt de données les dimensions vont se dresser, dans quelles tables de fait les estimateurs vont s'ajouter et comment la constellation va articuler les différents schémas en étoile obtenus.

1.1 Quel processus de modélisation dimensionnelle ?

La construction d'un premier entrepôt de données en une seule étape est une tâche quasi insurmontable. Le processus de production de résultats est constitué de l'ensemble des informations portant sur les formations boisées ou arborées à caractère forestier. Les protocoles s'adaptent suivant le type d'inventaire pris en compte. Toutes les données ne sont donc pas traitées de la même manière : sur chacun de ses types, les variables mesurées et estimées diffèrent. L'étude porte sur les domaines d'études suivants : les forêts de production, les peupleraies et les landes. La méthode de calcul des résultats portant sur les ligneux hors forêt n'étant pas finalisée, il n'en sera pas tenu compte.

	Dimensions communes				
Variables estimées	Point	Forêt – Bosquet - Peupleraies	Landes	Arbre	Souche
Surface	X	X	X		
Volume, Effectif, Surface terrière, Biomasse, …	X	X		X	
Volume, Effectif, Prélèvement, Circonférence	X	X			X

Tableau 4-5. Marché d'information de l'IFN modélisé dans l'entrepôt de données.

Selon le type d'inventaire choisi, les variables proposées seront plus ou moins nombreuses. Le tableau ci-dessus résume les cas envisagés. Par exemple, seules des données de surface seront disponibles pour des domaines d'étude pour lesquels les équipes de l'IFN ne couvrent pas ou partiellement leur étendue (territoire inventorié, agricole, sans végétation, eau). A l'inverse de nombreuses variables seront disponibles pour les « Forêt de production hors peupleraie ».

A ce stade de la modélisation, nous pouvons déjà établir que chaque domaine d'étude, par leurs natures d'opérations d'inventaire différentes, pourraient constituer des modèles différents. La figure 1 ci-dessous résume les modèles dimensionnels potentiels choisis dans notre marché d'informations forestier.

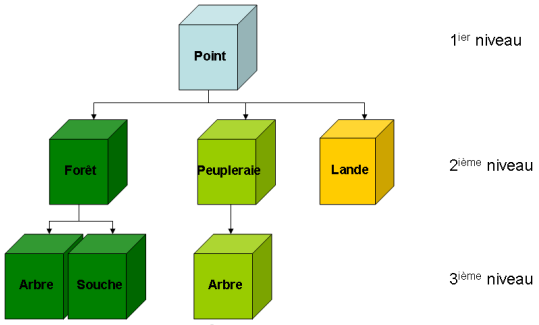


Figure 4-1. Modèles dimensionnels candidats.

Le modèle « Point » rassemblera toutes les informations de type reconnaissance (nature du point, statut juridique de la propriété, couverture du sol, taille de formation, …) puisque ces informations sont collectées sur l'ensemble des points visités. Les modèles « Forêt », « Peupleraie » et « Lande » contiennent toutes les informations levées au niveau du point. Enfin les modèles « Arbre Forêt », « Arbre Peuplier » et « Souche » couvrent l'ensemble des informations des arbres vivants et morts en forêt et peupleraie.

Le protocole des peupleraies est quasiment semblable à celui des forêts, cependant l'IFN ne pratique pas d'étude écologique et floristique sur les peupleraies car ces dernières sont entretenues. Dans notre version de modélisation, sont donc distingués deux modèles dimensionnels distincts. Suite à l'avis du directeur technique de l'IFN et dans un but d'optimisation, comme le protocole terrain le fait depuis peu, les données forêt et peuplier pourraient être rassemblées et ainsi ces deux modèles seraient fusionnés. Il suffira alors d'ajouter une modalité « hors contexte » pour les données qui n'existent pas (par exemple pour les lignes « forêt » avec la donnée « entretien de la peupleraie »).

Les modèles du deuxième et troisième niveau ne peuvent pas être fusionnés. En effet les mesures stockées dans les tables de faits de chacun de ces modèles ne sont pas du même grain. Cependant les modèles du troisième niveau ont la particularité d'hériter de toutes les dimensions du deuxième niveau. Mais les mesures d'un arbre ne peuvent pas être remontées au niveau des peuplements (modèles du deuxième niveau). Par exemple les mesures enregistrées dans la table de fait « Forêt » sont prises à l'intersection de toutes les dimensions contenant les informations relevées au niveau du point d'inventaire. En revanche les mesures enregistrées dans la table de fait « Arbre » sont prises à l'intersection de toutes les dimensions contenant les informations relevées au niveau du point et toutes les dimensions contenant les informations relevées au niveau de l'arbre.

Contrairement au cas précédent, la table de fait du modèle du premier niveau et les tables de fait des modèles du deuxième niveau définissent un même type de mesure : des surfaces. Il aurait été possible de les rassembler en un seul modèle. De plus toutes les dimensions du modèle point sont des dimensions des modèles « Forêt », « Peupleraie » et « Landes ».

Tout d'abord le volume d'informations n'est pas le même : dans le modèle « Point », toutes les lignes permettant de calculer des surfaces à partir des points de l'échantillon vont pouvoir être stockées, qu'ils aient été visités ou non. Ce modèle référence les points visités sur le terrain ou recopiés de la phase de photo-interprétation parce que non visités (confirmés sans visite terrain car trop couteux). En revanche les modèles du deuxième niveau, comporte uniquement les points visités, et ils sont de plus caractérisés par d'autres dimensions liées aux données relevées et à la nature du point. Si les modèles avaient été fusionnés (comme c'était envisagé dans le cas précédent : forêt et peupleraie), il aurait fallu procéder à l'ajout d'une dimension « Lever » avec deux membres supplémentaire « Non levé » et « Levé ».

Doit-on construire une table de fait rassemblant tous les points levés ou non et avec toutes les dimensions liées aux protocoles des différents domaines d'étude pour voir toutes les informations du point ou doit-on construire une table de faits séparée pour chaque domaine d'étude et pour l'opération de lever ?

Afin de faire ce choix les besoins d'analyse des utilisateurs devront être pris en compte. L'analyse va devenir très vite complexe ainsi que la présentation des données si l'option d'un seul modèle dimensionnel est retenue. De plus les analyses les plus courantes concernent un domaine d'étude avec ces particularités propres (écologie pour les forêts, cartographie pour les points de l'échantillon, ….)

Une autre considération que l'on peut prendre en compte c'est qu'aujourd'hui les mesures des faits de ces deux niveaux sont identiques (la nouvelle méthode débute, on ne calcule encore que la surface au niveau peuplement). Mais dans un futur proche, d'autres mesures vont s'ajouter comme par exemple la surface terrière moyenne dominante des arbres sur un point forêt, la hauteur des 50 plus gros arbres sur un point forêt, …. Ces nouvelles mesures n'ont pas de sens dans les autres domaines d'études.

Il ne convient donc pas de fusionner ces modèles dimensionnels. Au final on obtient des tables de faits distinctes pour les points de notre échantillon, les points forêt, les points peupleraies et les points landes. Cette décision s'impose naturellement parce que les données de chacun de ces modèles obéissent à un protocole propre et forment des dimensions spécifiques pour chaque domaine d'étude. L'approche de la table de faits unique aurait obligé à ajouter un membre « Hors Protocole » dans chacune des dimensions, et les requêtes d'analyse devenaient compliquées à réaliser (est-on bien sur une dimension valide ? les points sont-ils levés ?).

Au final, sept schémas en étoile ont été implémentés représentés dans la figure 4-1.

1.2 Constellation

Ce qui était aussi important dans cette première phase de modélisation c'est de bien avoir vu que de nombreuses dimensions allaient se partager. Par exemple dans les dimensions cartographiques, les dimensions « Propriété du point » ou « Localisation administrative du point » vont se joindre à chaque modèle dimensionnel envisagé.

Au moment de leur implémentation les dimensions communes devront bien avoir toutes la même similarité. Elles utiliseront les mêmes clés de dimensions, les mêmes noms de colonnes d'attributs les mêmes définitions d'attributs et les mêmes valeurs d'attributs, ce qui garantira des intitulés d'états et des regroupements d'information cohérents.

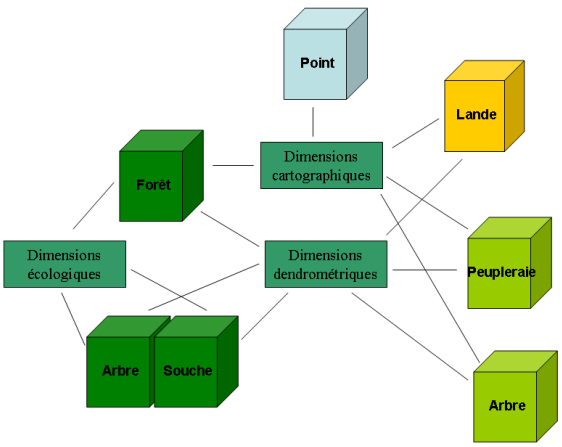


Figure 4-2. Modèle potentiel en constellation pour l'IFN

A ce stade de la modélisation, seul un aperçu des modèles potentiels de la constellation des inventaires forestiers a été présenté.

1.3 Déclaration de la granularité

La deuxième étape de la modélisation du processus consiste à définir le niveau de granularité des tables de faits. Cette étape est critique, car si un grain trop élevé est choisi, le modèle se limite à un nombre plus faible de dimensions ou à une information moins détaillée. Un modèle trop grossier est vite vulnérable à des demandes imprévues et inattendues des utilisateurs qui voudraient faire un forage plus profond. Cependant enregistrer les données cumulées joue un rôle important dans l'optimisation des performances. Au contraire, enregistrer les données les plus atomiques offre une grande souplesse : elles sont hautement dimensionnelles et sont prêtes à subir les assauts des demandes ad hoc des utilisateurs, c'est ce que qui est vivement recommandé.

Plus précisément il faut spécifier exactement ce que représente une ligne individuelle de chacune des tables de fait. Prenons par exemple le modèle « Forêt » dans lequel les mesures concernées sont le comptage des points et la surface. Il faut spécifier le niveau de détail de ces mesures, c'est-à-dire le plus petit niveau d'information que l'utilisateur voudrait. La mesure du comptage des points étant un peu particulière, focalisons-nous sur la mesure surface.

En plus des dimensions dendrométriques (DD), écologiques (DE) et cartographiques (DC) (détaillées plus tard), la dimension « Année » est une dimension très importante. A cause de cet axe toutes les mesures ne sont pas additives, mais semi-additives. Afin d'estimer les superficies forestières il faut effectuer au final une moyenne temporelle simple sur la période voulue. Par exemple si l'on cherche à estimer la surface en propriété privée des résineux en Aquitaine sur la période [2005-2009], il faut effectuer une moyenne des estimations de ces surfaces sur chacune des années, soit :

$$S_{[2005-2009]} = \frac{S_{2005} + S_{2006} + S_{2007} + S_{2008} + S_{2009}}{5}$$

Le calcul de la surface n'étant pas simple, rappelons son calcul défini au chapitre 3. Pour un sous-domaine d'étude f (par exemple une zone propriété privée des résineux en Aquitaine), $K = \bigcup_{k \in K} k$ l'ensemble des strates du territoire, la surface S_f du sous-domaine d'étude recherchée vaut pour une année :

$$S_f = \sum_{k \in K} \left(S_k * \frac{\sum_{i \in f \cap k} w_i}{\sum_{j \in k} w_j} \right)$$ Où w_i est le poids du point i de notre échantillon.

Le sous-domaine d'étude f correspondrait à une sous-coupe du modèle suivant tous les axes des dimensions dendrométriques (DD), écologiques (DE) et cartographiques (DC) et pour un membre de la dimension Année.

D'après l'équation précédente, le calcul de la surface dépend du découpage du territoire français en strate. Pour la plupart des utilisateurs, le choix de la stratification (opération de découpage en strate) est transparent (une par défaut), mais pour des experts avertis, ce choix est possible. En effet il convient mieux de choisir un découpage où la modalité cartographique choisie dans le slice ou dans le forage est présente dans la stratification. Par exemple, l'utilisateur qui désire connaître le volume en bois par région administrative, préfèrera la stratification « de phase 2 par département et types FAO » au lieu de celle « de phase 2 par région forestière et types FAO » (les régions administratives sont des regroupements de département ce qui n'ajoute pas d'erreur supplémentaire contrairement aux régions forestières qui chevauchent les régions administratives et engendrent alors des erreurs dans le calcul des estimations : estimation de la surface des départements au lieu d'utiliser leurs valeurs SIG).

La stratification dépend de l'année (car les données de photo-interprétation sont utilisées comme critère de stratification). En effet la définition des strates (le contour cartographiques des départements peut changer) et par conséquent leurs surfaces évoluent dans le temps. En reprenant la stratification « de phase 2 par département et types FAO », son découpage en strate peut être modifié entre l'année 2006 et 2007, par un changement de combinaisons des critères, par une modification de surface ou dans le nombre de strates obtenues.

Il n'est donc pas possible de passer outre la dimension de la stratification, surtout si l'on veut garder ce critère comme axe de comparaison (nous obtenons des estimations différentes de la surface à chaque stratification choisie).

En poussant le forage au plus profond, le grain le plus fin que l'on pourrait obtenir serait en ajoutant la dimension supplémentaire du point. A priori, les estimations obtenues au niveau du point n'ont pas vraiment de sens : qu'est-ce qu'une surface forêt ventilée par les identifiants des points ?

Si la dimension du point se présentait cela apporterait des informations supplémentaires, comme ses coordonnées, la date du levé, l'équipe responsable des relevés, …. Toutes ces informations correspondent à un processus différent : le processus de collecte des données d'inventaire. Les indicateurs s'identifieraient alors à des analyses de suivi (nombre de points visités par telle équipe, les régions en retard sur le planning prévisionnel, …), ce qui conviendrait très bien à une mise en place d'un tableau de bord spatial. Mais cette approche serait le sujet d'une autre étude. Dans le calcul des estimations que nous voudrions faire, la dimension « Point » n'a, aujourd'hui, pas d'intérêt.

Toutefois, cet ajout aurait cependant un grand avantage. Ayant modélisé notre modèle au niveau atomique le plus fin, l'ajout d'une nouvelle donnée, ou la modification d'une définition (nouvelle discrétisation des classes de diamètre de l'arbre par exemple) ne présenterait aucune nouvelle modélisation du modèle, et ainsi aucun nouveau chargement.

Cette dimension n'a pas été ajoutée, la granularité au niveau de la strate semblait suffisante afin de produire des résultats.

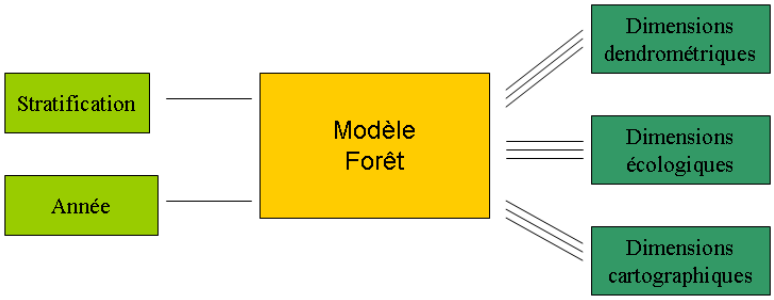


Figure 4-3. Granularité du modèle dimensionnel "Forêt".

En résumé, la granularité du modèle forêt est défini par le croisement des dimensions « Année », « Stratification » et de l'ensemble des dimensions de type dendrométrique, cartographique et écologique comme le résume la figure 4-3.

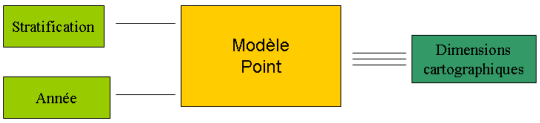


Figure 4-4. Granularité du modèle "Point".

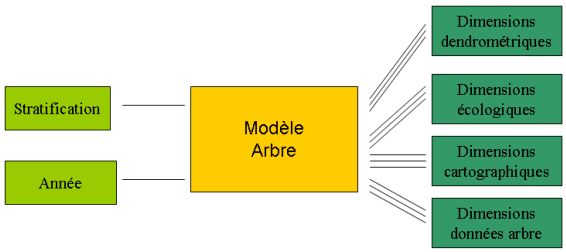


Figure 4-5. Granularité du modèle "Arbre".

Pour les autres modèles le même principe a été gardé (figures 4-4 et 4-5).

1.4 Hétérogénéité

Bien que la nouvelle méthode mise en place à l'Inventaire Forestier soit assez récente, il ne faut pas négliger les problèmes d'intégration des données recueillies à différents moments d'observation. Au fil du temps de nouveaux protocoles ont été mis en œuvre créant potentiellement une hétérogénéité sémantique, temporelle et spatiale. Chaque année le protocole peut évoluer ce qui se traduit notamment par des définitions de variables qui changent (cubage des arbres, détermination de l'âge moyen, calcul de la surface terrière, …), certaines données ne peuvent plus être relevées, d'autres données sont nouvellement prises. Le découpage des zones cartographiques évoluent aussi, à fréquence fixe des nouvelles cartes sont générées.

Les modèles dimensionnels sont fondés sur une organisation des données dont les dimensions sont statiques et seules les tables de fait reflètent l'aspect dynamique. Ce paragraphe tente de voir comment résoudre cette problématique d'intégration de ces informations évolutives dans ces dimensions.

Tout d'abord, l'intégration de nouvelles données n'engendre pas de difficultés si la granularité a bien été définie, seul un nouveau modèle est constitué. Si la granularité est trop grande, une opération de rechargement est alors nécessaire. Par exemple en ajoutant une nouvelle donnée issue d'une composition de données déjà existantes (l'exploitabilité très difficile n'est plus définie pour les pentes supérieures à 30% mais supérieures à 50%), le modèle est ajusté pour tenir compte de cette donnée, mais aucun chargement n'est nécessaire. En revanche si une donnée est nouvellement mesurée, alors il faut envisager un nouveau chargement.

Grâce à la base des métadonnées de l'IFN, on sait pour chaque donnée à quelles campagnes consécutives elle a été prise. Par conséquent pour les données qui ne sont plus relevées (l'accessibilité, la portance du sol, …) ou qui sont nouvellement relevées (l'aspérité, structure forestière, …) il faut ajouter un nouveau membre dans la dimension « Hors limite temporelle ».

En ce qui concerne les données qui changent de définition, mais pas de plages de valeur, comme les données discrètes, il n'y a pas d'impact pour l'intégration dans l'entrepôt de données. C'est une information à communiquer aux utilisateurs (mode de calcul de la donnée). Par exemple le calcul de l'exploitabilité rentre dans cette catégorie. Lors des trois premières campagnes le calcul de l'exploitabilité repose sur la combinaison de quatre données relevées sur le terrain : l'accessibilité du point, la distance de débardage, la portance du terrain et la pente du sol. Ce calcul est simplifié pour les peupleraies. Lors des campagnes suivantes, le calcul de l'exploitabilité repose sur la combinaison de 5 données levées sur le terrain : l'itinéraire de débardage, la distance de débardage, la portance du terrain, l'indicateur de pente de débardage et l'indicateur d'aspérité. Au fil du temps les tranches de valeur de cette donnée restent « facile », « moyenne », « difficile » et « très difficile ». La dimension « Exploitabilité » de l'entrepôt reste donc identique.

Pour les valeurs des données qui évoluent dans le temps, une solution consisterait à faire évoluer les dimensions elles-mêmes en ramenant les données à une version unifiée. Cela apporterait une uniformité et une homogénéité des données de ces nouvelles dimensions, en revanche il peut s'avérer que l'on perde une quantité d'informations non négligeables. Pour la dimension « Essence », par exemple, les classes de valeur avant 2006 de l'essence du chêne étaient toutes rassemblées sous la valeur « Chêne indifférencié », mais depuis les membres spécifiques sont détaillés jusqu'à sept valeurs (Chêne vert, Chêne pédonculé, Chêne liège, ….). La première solution consisterait à créer des modèles pour les niveaux comparables : un pour 2005 et un second pour 2006 et plus. Cette solution limite les capacités d'analyse et ne convient donc pas. La deuxième solution consiste à intégrer les données détaillées et agrégées dans le même modèle et en ajoutant un membre « Non valide ». Ainsi les évolutions temporelles des dimensions thématiques sont conservées. Cette solution reste complexe pour les analyses. Les estimateurs sont des moyennes temporelles, si une donnée comme l'essence principale est choisie dans une ventilation, les résultats peuvent être difficiles à interpréter. Afin d'avoir une analyse simple un seul modèle a été implémenté mais uniquement avec des dimensions

agrégées sur les années des campagnes de l'entrepôt de données (2005 à 2009). Dans notre cas, la dimension « Essence » ne comporte que la modalité « Chêne indifférencié ».

Pour ce qui est de l'hétérogénéité spatiale, l'usage qui en est fait dans l'entrepôt de données n'apporte pas encore de problèmes fonctionnels. Les informations spatiales de l'entrepôt qui évoluent ne concernent que la propriété du contour géographique des différents membres des dimensions cartographiques. Cette propriété n'a aujourd'hui qu'un usage de représentation (abordée dans le chapitre suivant). Si on choisit par exemple de stocker le contour géographique du département « Loiret » de 2005 dans la dimension « Localisation administrative », l'affichage des résultats des superficies forestières de 2007 du Loiret seront dessinées sur ce contour, ce qui n'est qu'une illustration de l'information.

En revanche, le problème existe pour les régions forestières dont la définition va changer (nouvelle méthode de conception). Dans ce cas on ne parlera plus de région forestière mais de sylvo-éco-région. Mais ce problème, à notre niveau, est identique à l'ajout d'une nouvelle dimension dans l'entrepôt de données, il faut redéfinir un modèle et un nouveau chargement.

1.5 Choix des dimensions

Maintenant que les modèles et les tables de fait sont identifiés, il reste à les enrichir d'un ensemble fourni de dimensions représentant toutes les descriptions susceptibles de prendre des valeurs particulières dans le contexte de chaque mesure. Dans les modèles définis précédemment différents type de dimensions sont répertoriées. La dimension « Stratification-Année » qui a un rôle très particulier, les dimensions classiques textuelles qui regroupent les informations dendrométriques, écologiques et cartographiques, et les dimensions spatiales qui comportent un attribut spécial de géométrie.

1.5.1 Dimensions thématiques

La dimension est la première information parmi les mesures et les attributs qu'un analyste explorera. Commençons par décrire les dimensions les plus classiques et les plus nombreuses : celles qui comportent des listes d'attributs textuels distincts et qui vont étoffer chaque table de dimension. Elles sont répertoriées dans quatre catégories où les données relevées présentent des caractéristiques homogènes et variées :

- Les dimensions dendrométriques qui vont caractérisées toutes les informations des peuplements. Elles sont liées au point d'inventaire. Elles peuvent être spécifiques à un domaine d'étude ou communes à quelques domaines d'études. Il est donc important d'identifier sur quels objets elles portent une caractéristique afin de pouvoir les partager ou non (si une dimension comporte un attribut spécifique à un unique domaine d'étude, la dimension ne pourra pas alors se joindre aux autres modèles).

- Les dimensions « données arbre » comportent aussi des informations dendrométriques mais au niveau arbre ou souche (bois mort). De la même façon, certaines données sont partagées par les arbres forêt et peupliers.

- Les dimensions écologiques vont caractérisées des données écologiques au niveau des points forêt.

- Les dimensions cartographiques qui vont caractériser toutes les informations dites de reconnaissance du point. Elles sont issues du service cartographie, d'où leur nom, mais ne présente pas dans notre étude, un intérêt spatial. Ces données sont importées soit à partir d'un croisement cartographique réalisé en amont du chargement de l'entrepôt, soit à partir de la base des données de photo-interprétation.

Les données que composent les bases de données d'exploitation de l'IFN sont de différentes natures : elles peuvent être brutes (issues directement de relevé), composées ou calculées à partir des données brutes, discrétisées ou regroupées. Cependant toutes les données ont un petit nombre fini de valeurs (jamais plus d'une centaine), aucune donnée n'est textuelle.

Définir les dimensions consiste à rassembler des éléments de la base de production de même nature tels que les données calculées sont déduites des attributs de cette même dimension. Les données corrélées les plus faciles à déterminer sont les données regroupées (information stockée dans la base des métadonnées de l'IFN). Pour ce qui est des données composées, il manquait l'information des données paramètres qui permettent de les calculer.

La première version de la modélisation des dimensions repose sur une architecture « Mille-pattes » c'est à dire avec un grand nombre de dimensions ce qui n'est pas très optimum comme solution. Il paraît opportun après coup qu'indexer l'énorme clé de la table de faits est difficile et que les nombreuses jointures allait réduire la commodité d'utilisation et la performance. D'après Ralph Kimball (inventaire en 1982 du concept d'entrepôt de données), il faut rester en-dessous de 15 dimensions et au-delà de 25 dimensions il faut chercher des moyens de combiner des dimensions corrélées en une seule dimension. Des attributs parfaitement corrélés (tels que des regroupements d'attributs, des attributs avec une corrélation statistique raisonnable) doivent faire partie d'une même dimension. On obtient un résultat satisfaisant quand l'unique dimension résultante est nettement plus petite que le produit cartésien des dimensions combinées. Il est même préconisé de créer des dimensions fourre-tout avec une liste de données restantes pouvant prendre un nombre restreint de valeurs. Il faut avant tout essayer de limiter la taille de la table de faits en augmentant au minimum la taille de la ligne.

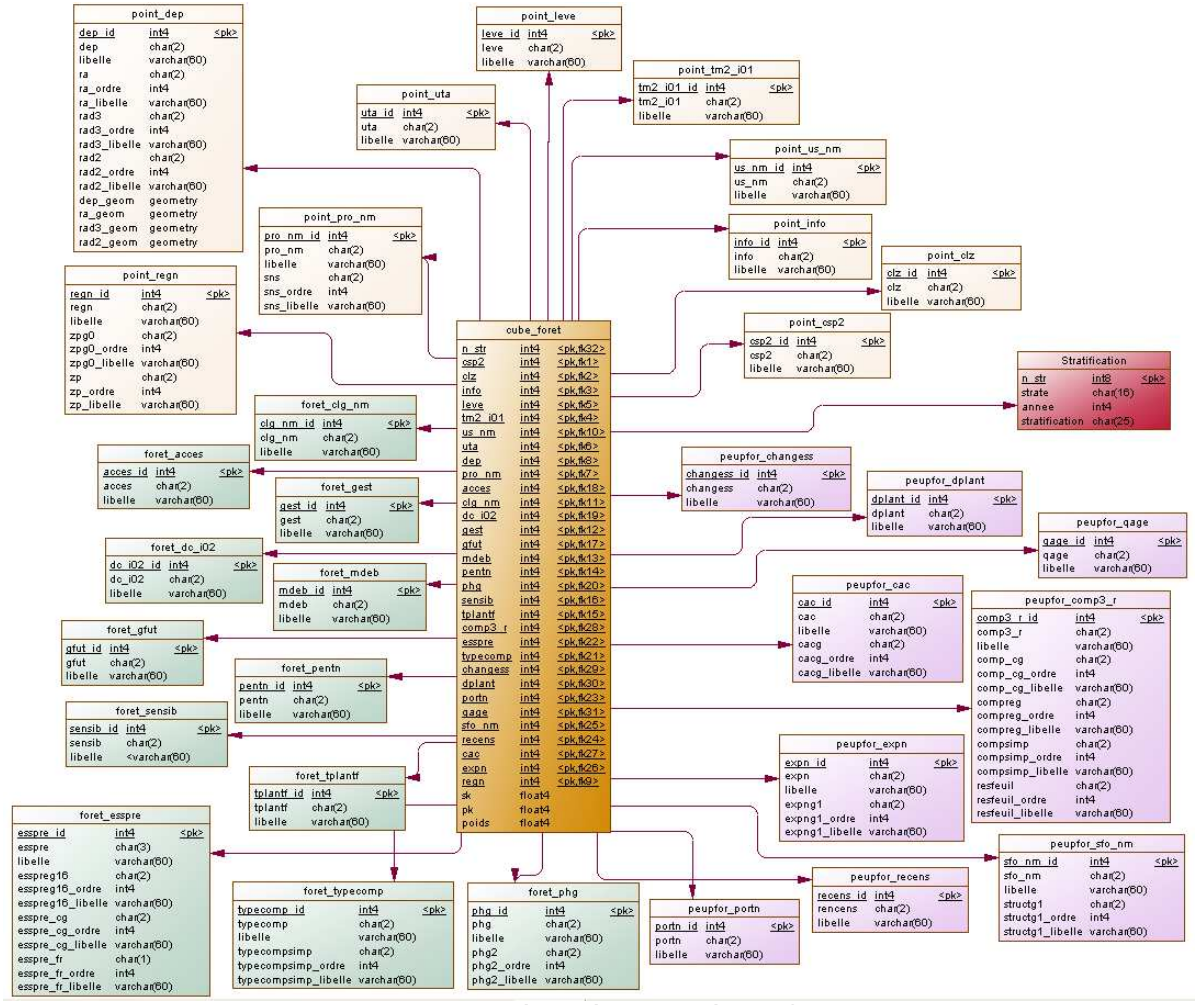


Figure 4-6. Schéma en étoile du modèle "Forêt"

N'ayant pas vu tout de suite cette méthode, le modèle dimensionnel obtenu correspond à une dimension pour chacune des données brutes, calculées et discrétisées. Et pour chacune d'elles leurs unités regroupées ont constituées différents niveaux des hiérarchies. La figure résume l'ensemble des dimensions du modèle « Forêt ». Les performances, abordées dans le dernier paragraphe, sont toutefois correctes. Remarquons tout de même que beaucoup de dimensions comportent une ou plusieurs hiérarchies d'au moins un niveau voir plus (Essence principale, Composition du peuplement recensable, Exploitabilité, Structure forestière, Localisation, Propriété, Type d'humus…)

Afin d'identifier rapidement les dimensions partagées entre les différents modèles, le nom des tables a été choisi en conséquence. Dans notre exemple, un nom commençant par « peupfor_ » correspond à une dimension du modèle forêt et du modèle peupleraie.

1.5.2 Dimensions spatiales

Les opérations d'inventaire sont réalisées sur l'ensemble du territoire français. Historiquement et pour des besoins liés à sa fonction d'établissement public, les résultats pour l'ensemble du pays étaient ventilés par le département ou par les régions administratives. Les régions forestières (divisions territoriales présentant des points de vue des conditions au sol, de climat et de croissance des peuplements forestiers) ont été créées et des besoins de connaître des résultats pour celles-ci s'est ensuite fait ressentir. L'information spatiale est donc un axe d'analyse à l'IFN.

Dans les solutions OLAP traditionnelles la dimension spatiale est traitée comme une dimension descriptive avec une représentation hiérarchique telle que vue dans le chapitre précédent sans aucun lien vers une représentation cartographique.

Une dimension spatiale est caractérisée par la présence de l'attribut géométrique dans les membres des différents niveaux. Dans nos modèles les hiérarchies spatiales sont définies en attribuant à chaque niveau une définition géométrique : le contour géographique. Elles représentent différentes granularités de l'information géographique, chaque niveau géographique représentant une information géographique différente. Comme les dimensions classiques d'un modèle dimensionnel, les dimensions spatiales sont agrégatives d'un niveau hiérarchique à un autre.

- « Dimension Administrative » permet l'analyser les faits suivant une vision purement administrative. Cette dimension se décomposerait en cinq niveaux hiérarchiques :
  - Niveau « Pays » : contenant le polygone France entière
  - Niveau « 5 inter-régions » : contenant les polygones des regroupements des régions administratives en 5 grandes régions.
  - Niveau « 9 inter-régions » : contenant les polygones des regroupements des régions administratives en 9 grandes régions.
  - Niveau « Région administrative » : contenant les polygones de chaque région administrative.
  - Niveau « Département » : contenant les polygones de chaque département.

- « Dimension Ecologique » dont la structuration en zone écologique homogène permet des analyses pertinentes dans le cadre de la gestion forestière. Cette dimension se décompose en quatre niveaux hiérarchiques :

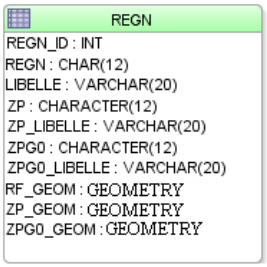


Figure 4-7. Dimension Spatiale Localisation selon la hiérarchie Ecologique

- Niveau « Pays » : contenant le polygone France Entière
- Niveau « Zone phytogéographique » : contenant les polygones des 3 zones phytogéographique (Montagnes, Plaines-Collines et Région Méditerranéenne).
- Niveau « Zone Ecologique » : contenant les polygones des grandes régions écologiques.
- Niveau « Région forestière nationale » : contenant les 309 régions forestières.

Dans certains cas il est plus judicieux de suivre les indicateurs d'analyse selon un découpage administratif mais aussi dans d'autres cas plus selon un découpage géographique écologique. Ainsi pour la même dimension « Localisation » on pourrait avoir deux hiérarchies possibles. On parle alors de « Hiérarchies alternatives » ; l'utilisateur choisira soit l'une soit l'autre pour effectuer sa propre analyse. Ainsi on a bien deux hiérarchies pour la même dimension spatiale et dont chacune des deux fait appel à des objets géographiques différents.

Actuellement on s'autorise dans les tableaux à croiser ces deux dimensions c'est pourquoi les dimensions sont séparées. Toutefois le problème se posera dans l'application web OLAP cartographique que nous détaillerons dans le prochain chapitre.

1.5.3 Dimension stratification et année

Cette dimension est très spécifique à l'Inventaire Forestier. Elle pourrait s'apparentée à la dimension date par son caractère temporel et aussi par le fait que c'est la première dimension dans l'ordre de tri de la base de données, de sorte que les chargements successifs de données sont liés à une campagne d'inventaire et une stratification.

Comme nous l'avons vu dans le paragraphe de définition de la granularité, cette dimension est non-additive pour toutes les mesures hormis celle du comptage des points.

Chaque ligne identifie une strate pour une année et pour une stratification.

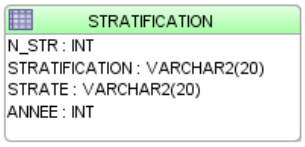


Figure 4-8. Schéma de la dimension « Stratification »

n_str integer	strate character(16)	annee integer	stratification character varying(100)
1	01.2F50	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
2	01.2SV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
3	01.3AV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
4	01.3EAU	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
5	01.3F1B	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
6	01.3F3B	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
7	01.3F50	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
8	01.3LHF+BOSQ	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
9	01.3SV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
10	02.1AV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
11	02.1F50	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
12	02.1SV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
13	02.3AV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
14	02.3EAU	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
15	02.3F3B	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
16	02.3F50	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
17	02.3LHF+BOSQ	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
18	02.3SV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
19	03.1AV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
20	03.1F50	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
21	03.1SV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
22	03.3AV	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1
23	03.3EAU	2005	Stratification Ph2 par DEP, types FAO, PROpriete et Ph1

Figure 4-9. Détail de la table de dimension "Stratification"

La colonne « Année » contient l'année de la campagne et donc l'année de l'échantillon. Cette colonne est très importante car c'est un paramètre de la fonction de calcul des estimateurs. Elle est de type entier plutôt que de type date afin d'optimiser les requêtes.

La dimension « Stratification » va définir les différents découpages. Actuellement il n'existe qu'une seule valeur, ce sera donc celle par défaut.

Pour chaque année et une stratification, on a un jeu de strates complet, ce qui veut dire que la somme des surfaces des strates correspond à la surface de la France métropolitaine. Chaque point de notre échantillon appartient à une seule strate et chaque strate contient au moins un point. Ces propriétés vérifiées garantissent la validité des calculs des estimateurs. Dans la première version d'implémentation des cubes (Cf. paragraphe 2) la dimension strate dans les schémas logiques n'a pas encore d'utilité, elle ne sert actuellement qu'à définir la granularité dans le modèle physique.

Cette dimension est vouée à devenir très volumineuse avec l'incrémentation des campagnes d'inventaire : une stratification pour une année apporte en moyenne 850 lignes. Nous verrons dans le paragraphe 3 comment optimiser les requêtes en créant des tables agrégats.

1.6 Identifier les faits

Ultime étape à la modélisation complète de la constellation de l'IFN, il faut maintenant identifier les mesures qui apparaîtront dans les tables de faits. Pour les modèles dimensionnels du premier et du second niveau, comme pour l'instant seule la surface est estimée, il faut stocker la somme des poids des points déterminés par l'intersection de toutes les dimensions : « Année », « Stratification », « Localisation Administrative », « Localisation Ecologique », « Propriété », … comme indiqué à la figure 4-6. Dans cette première modélisation les données surface de la strate (Sk) et poids de la strate sont stockées dans la dimension « Stratification », car ils apparaissent plus comme des attributs de la dimension.

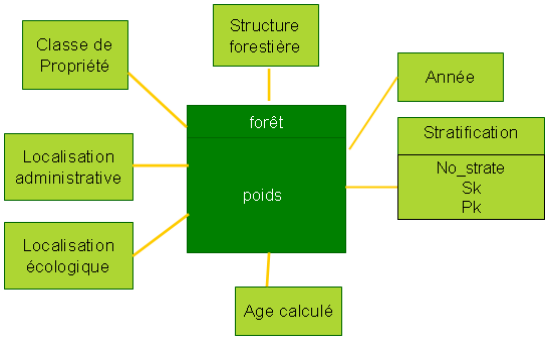


Figure 4-10. Faits mesurés du modèle "Forêt"

Seulement la surface du sous-domaine d'étude se calcule par la somme des surfaces des strates (Sk) multiplié par la proportion du sous-domaine dans la strate (soit le rapport de la somme des poids des points de ce sous-domaine divisé par la somme des poids des points de la strate (Pk)). Les variables Sk et Pk sont donc nécessaires au calcul, ce qui veut dire que ce ne sont pas des attributs mais des faits. Leur particularité est qu'ils ne sont ni additifs, ni semi-additifs parce qu'on ne peut les agréger sur aucune des dimensions.

Finalement le modèle dimensionnel « Forêt » se présente comme dans la figure 4-7.

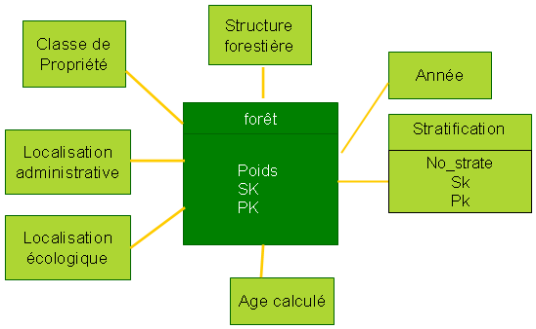


Figure 4-11. Faits mesurés du modèle "Forêt"

Pour les modèles du troisième niveau, il faut stocker en plus la moyenne des variables quantitatives dans le sous-domaine et la strate dont on veut calculer les estimations (volume sur pied, surface terrière, effectif, biomasse, prélèvement, circonférence, …).

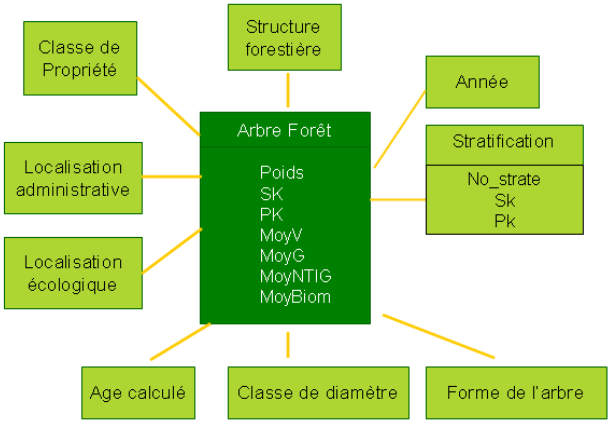


Figure 4-12. Faits mesurés du modèle "Arbre"

Par exemple la figure 4-12 illustre les dimensions et la table de fait du modèle dimensionnel « Arbre Forêt ».

2 Intégration des données

L'activité qui suit celle de la modélisation est le développement de l'application ETC afin d'alimenter notre entrepôt de données. Nous décrirons comment nous avons procédé afin d'extraire les données en provenance des bases de données opérationnelles de l'IFN, puis comment nous les avons structurées, sélectionnées et agrégées au sein de notre entrepôt.

2.1 Préparation des tables dimensionnelles

La première phase de l'ETL est la phase d'extraction qui consiste à préparer les données à partir des données brutes des applications opérationnelles et de les adapter au modèle dimensionnel défini dans la zone de préparation. La transformation est l'action de combiner les données, régler la qualité, identifier les mises à jour, gérer les clés artificielles, construire les agrégats et traiter les erreurs. Le dernier processus, celui de chargement, est simple à exécuter dès que les données ont été correctement préparées.

L'étape la plus longue a été de concevoir les tables dimensionnelles. En effet dans cette démarche de nombreux cas particulier demandaient des traitements spéciaux.

La question s'est posée sur le choix d'un ETC. Après avoir installé et testé Talend Open Studio qui est un client lourd open source très robuste et complet, la suite n'était pas très adaptée au système de l'Inventaire, qui a déjà sa propre gestion documentaire (la base de données MetaIFN présentée au chapitre 3). Une extension du stockage des métadonnées dans une base dédiée à l'entrepôt complétée d'un script de chargement, écrit en java, s'appuyant sur ces nouvelles métadonnées ont été implémentés. Le diagramme de classe de la gestion des métadonnées de l'entrepôt est représenté en figure 4-13 ci-dessous.

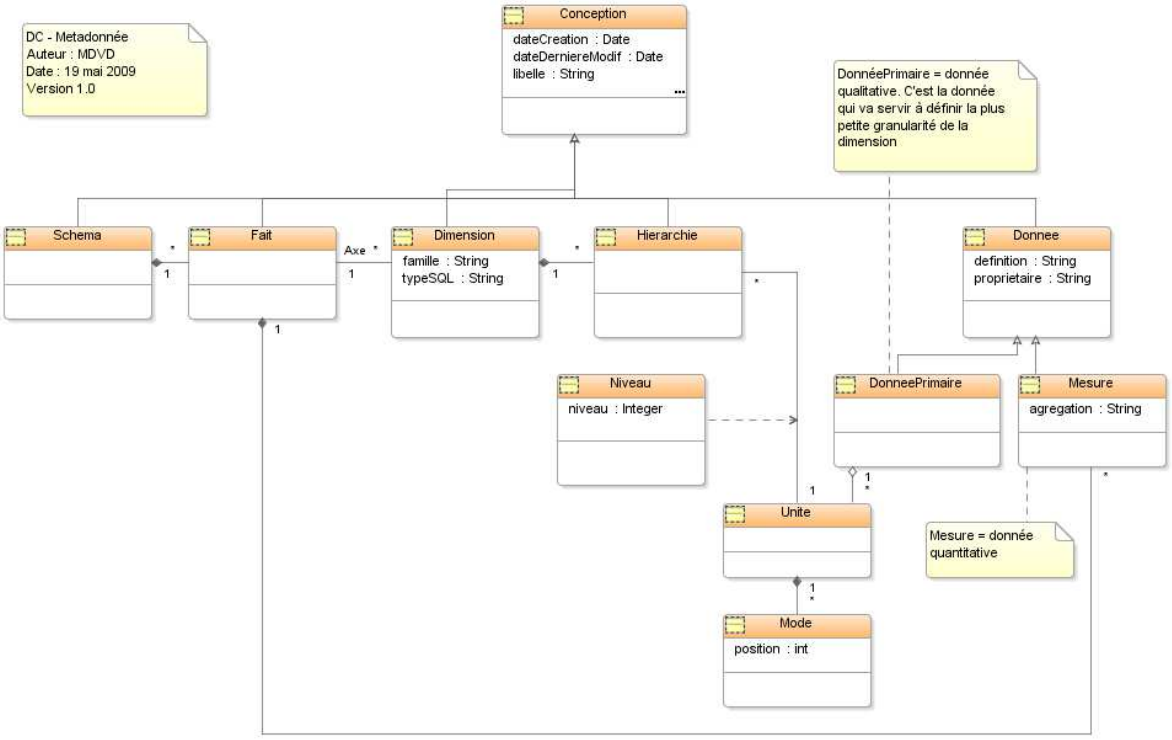


Figure 4-13. DC Implémentation

Les métadonnées de l'IFN permettent de définir rapidement quelles sont les données qualitatives potentielles vouées à devenir des dimensions et quelles sont les données quantitatives vouées à devenir des mesures. Comme nous l'avons expliqué un peu plus haut, seules les données dites regroupées vont devenir des attributs des dimensions, elles ne sont donc pas enregistrées dans la table `Dimension`. Toutes les autres données candidates sont donc des dimensions. Afin de définir les hiérarchies la tâche s'annonçait plus ardue. En effet la table de la base des métadonnées de l'IFN regroupe bien toutes les unités mais sans la notion de niveau et donc sans la notion de hiérarchie. Puis toutes les unités ne regroupent pas forcément l'unité de niveau inférieur. Toute cette phase a été réalisée manuellement. La liste des modalités de chaque unité est chargée sans modification.

Une fois cette base de métadonnées renseignée, la création des tables de faits et des dimensions, ainsi que le remplissage de ces dernières ont découlé immédiatement. Pour chaque dimension, hormis la dimension « Stratification », il faut encore ajouter deux valeurs possibles :

- Une ligne pour les valeurs à NULL en base
- Une ligne pour les dimensions qui ne sont pas valides pour cette campagne.

Ce qui donne, par exemple pour une dimension simple Altitude :

_clz [PK] integer	clz_clz character(10)	clz_clz_libelle character varying(100)
6	1	0 - 200 m
3	2	200 - 400 m
1	3	400 - 600 m
7	4	600 - 1000 m
4	5	1000 - 1400 m
2	6	PLUS DE 1400 m
5	xx	NULL
8	yy	HORS LIMITE TEMPORELLE

Figure 4-14. Données et attributs de la dimension "CLZ"

Les clés des dimensions sont artificielles, c'est-à-dire elles ont aucune signification et ne proviennent pas des bases de production. Elles sont affectées séquentiellement lors du remplissage de la dimension. Cela permet de garder une autonomie et d'être non dépendant des règles de gestion du

système opérationnel. Les clés artificielles offrent également de bonnes performances par leur nature entière.

La table dimensionnel de la stratification est une recopie d'une table qui renseigne les mêmes informations dans l'environnement d'exploitation de l'IFN. Contrairement aux autres dimensions, c'est une dimension qui évolue dans le temps. A chaque nouvelle campagne d'inventaire, il faut ajouter la nouvelle année et les nouvelles stratifications.

En dernier lieu, il restait à charger les attributs géométriques des dimensions spatiales. A partir des fichiers Shape, il suffit d'exécuter le script *shp2pgsql* dans des tables temporaires, puis de mettre à jour la dimension avec la clé du département ou de la région forestière. Les attributs géométriques des données regroupées se calculent en utilisant la fonction SQL : `ST_UNION`.

2.2 Chargement des tables de fait

Puisque le choix de la granularité n'a pas été prise au niveau du point mais au niveau de la strate et du domaine d'étude, les mesures sont agrégées dans la table de fait. Il faut donc effectuer des pré-calculs avant de charger les données. En créant des tables temporaires, on peut y ajouter les lignes pour chaque sous-domaine, chaque strate et chaque point des poids et des variables. Puis on peut agréger sur ces points et obtenir la somme des poids et les moyennes des variables par strate et par sous-domaine. Ces résultats agrégés ne sont plus qu'à charger dans les tables de faits.

De plus tous les points n'ont pas été chargés. L'avantage des entrepôts de données est qu'il comporte exactement les données dont les utilisateurs ont des besoins. Dans les bases de production, des points de tests y figurent ainsi que des points inventoriés que l'on ne doit pas prendre en compte (par exemple pour les deux premières campagnes, on ne prend pas les points dont la couverture porte sur « Autre Bosquet » et « Autre Foret »).

Les performances de la phase d'ETC sont acceptables (environ 5 heures), et les transactions ne dépassent pas l'espace de journalisation (cinq campagnes d'inventaire : 260 000 points et 350 000 arbres).

3 Analyse OLAP

Le modèle dimensionnel défini et intégré dans le système de base de données permet de fournir des données agrégées via des dimensions. Afin de réaliser des analyses OLAP, il nous faut définir le schéma, soit des cubes, des hiérarchies, des membres et la mise en correspondance qui transforme le modèle logique en un modèle physique. Nous commencerons par présenter dans ce paragraphe le serveur Mondrian qui a permis de mettre en place l'analyse OLAP, puis le modèle logique implémenté sera détaillé.

3.1 Base de données multidimensionnelles

Nous détaillons brièvement dans ce paragraphe les principaux concepts de base du modèle logique et leurs implémentations dans Mondrian.

3.1.1 Serveur Mondrian

Mondrian, comme nous l'avons vu dans le chapitre 2, est un serveur décisionnel libre écrit en Java. Mondrian fournit toutes les fonctionnalités les plus importantes d'un serveur OLAP, comme les hiérarchies multiples, les propriétés de niveaux des dimensions, les fonctions définies par l'utilisateur, les membres calculés, les cubes virtuels, le langage MDX, ….

Le modèle de la base de données multidimensionnelles, appelé aussi schéma, est défini à travers un fichier XML. Les principaux éléments d'un schéma sont : les cubes, les mesures et les

dimensions. Il représente aussi le mapping du modèle de l'application multidimensionnelle vers les tables stockées dans le SGBD. Des extraits de ce fichier sont fournis en annexe 1.

L'élément racine est donc le schéma attribué de son nom : <Schema name="DWIFN">.

3.1.2 Cube

Un cube est une collection de mesures et de dimensions qui ont la table de fait en commun. La table de fait contient les colonnes à partir desquelles les mesures sont calculées et des références vers les tables contenant les dimensions. Dans cet élément racine, les cubes sont définis ainsi :

```
<Cube name="Point" defaultMeasure="nb_pointsP">
   <Table name="cube_point" />
    ...
</Cube>
```

La table de fait est définie en utilisant le tag `<Table>`.

Le langage MDX permet d'utiliser plusieurs mesures par cube (il suffit d'utiliser la dimension `Measure`) ou même aucune. L'agrégation ne s'opère pas sur toutes les mesures mais seulement sur une seule définie par défaut dans le schéma modélisant le cube (`defaultMeasure`). Pour changer la mesure par défaut dans une requête MDX, il suffit de mentionner celle qui intéresse dans la clause where.

3.1.3 Mesures

Le cube « Point » définit une mesure « Nombre de points » :

```
<Measure name="nb_pointsP" column="point_nb" aggregator="sum" visible="true" />
```

Chaque mesure a un nom, une colonne de correspondance dans la table de fait et un opérateur d'agrégation. Pour calculer le nombre de points total on utilise l'agrégation « sum ».

Il existe deux sortes de mesures : les mesures non calculées et les mesures calculées.

```
<Measure name="Surf" aggregator="sum" visible="false">
    <MeasureExpression>
        <SQL dialect="generic">
        cube_point.point_poids * cube_point.point_sk / cube_point.point_pk / 1E4
         </SQL>
    </MeasureExpression>
</Measure>
```

Par exemple la mesure « Surf » définissant une surface sur une année et une stratification se définit non pas à partir d'une colonne en base mais par une expression SQL qui permet de calculer ses valeurs.

Cette mesure a comme attribut d'être invisible pour les applications utilisateurs comme JPivot. En réalité ce calcul n'est qu'une étape intermédiaire aux calculs plus complexes des estimateurs que nous verrons un peu plus tard.

Certaines mesures peuvent aussi se calculer à partir d'autres mesures. Par exemple le volume de Bois d'œuvre est la somme du volume de qualité 1 et du volume de qualité 2. Ce qui donne :

```
<CalculatedMember name="Volume de bois d'oeuvre" dimension="Measures">
        <Formula>[Measures].[VQ1] + [Measures].[VQ2]</Formula>
        <CalculatedMemberProperty name="FORMAT_STRING" value="#,###.00"/>
  </CalculatedMember>
```

3.1.4 Dimensions

Une dimension est un attribut, ou un ensemble d'attributs, à travers lesquels sont observées les mesures. Ajoutons ou rappelons quelques définitions :

- Un membre est un point dans une dimension déterminé par les valeurs d'attributs de cette dimension. « Privé », « Occulté », « Communal » et « Domanial » sont tous les membres de la hiérarchie « Propriété ».
- Une hiérarchie est un ensemble de membres organisés selon une structure appropriée à l'analyse. Par exemple les départements peuvent être regroupées par régions administratives qui peuvent par des inter-régions. Les mesures sont agrégées pour chaque niveau de la hiérarchie. La superficie forestière de la France est calculée à partir de la superficie forestière des inter-régions.
- Un niveau est une collection de membres qui ont la même distance de la racine à la hiérarchie.
- Une dimension est une collection de hiérarchies selon laquelle les faits sont observés.

```
<Dimension name="Propriété" foreignKey="pro_nm">
    <Hierarchy hasAll="true" allMemberName="total" primaryKey="pro_nm_id">
        <Table name="point_pro_nm"/>
        <Level name="Statut forestier" column="sns_libelle" uniqueMembers="true">
            <Property name="symbole" column="sns_libelle" />
        </Level>
        <Level name="Propriété" column="libelle" uniqueMembers="true"/>
    </Hierarchy>
</Dimension>
```

Ici la dimension « Propriété » est constituée d'une seule hiérarchie qui a deux niveaux. La dimension prend ses valeurs à partir de la colonne PRO_NM_ID de la table POINT_PRO_NM.

Une dimension est jointe à un cube à l'aide de deux colonnes, celle de la table de fait et celle de la dimension. Chaque dimension avec l'attribut `foreignKey` indique le nom de la colonne de la table de fait.

La clé du niveau d'une hiérarchie est définie par l'attribut `column`. Elle doit correspondre à une colonne de la table dimension.

L'attribut `uniqueMembers` est utilisé pour optimiser les commandes SQL. Si vous savez que les valeurs d'un niveau d'une table de dimension sont uniques sur toutes les valeurs d'un niveau parent, il faut alors mettre l'attribut `uniqueMembers="true"` et `"false"` sinon. Dans notre schéma, nous n'avons pas de cas où un membre n'est pas unique par rapport à son niveau supérieur. Dans le niveau le plus haut de la hiérarchie, l'attribut uniqueMembers est toujours à true car il n'y a aucun niveau parent.

Par défaut, toute hiérarchie contient un niveau haut appelé « All », qui contient un seul membre appelé « (All {hierarchieName}) ». Ce membre est le parent de tous les autres membres de la hiérarchie et représente l'agrégation totale sur cette dimension. C'est le membre par défaut de la hiérarchie. Il est le membre utilisé pour calculer les valeurs d'une cellule lorsque cette hiérarchie n'est pas incluse dans un axe ou un « slicer ».

Si l'élement « Hierarchie » a `hasAll="false"`, le niveau 'All' est supprimé, le membre par défaut d'une telle dimension est le premier membre par défaut.

```
<Dimension name="Année" foreignKey="n_str">
    <Hierarchy hasAll="false" primaryKey="n_str" defaultMember="[Année].[2008]">
        <Table name="strate"/>
         <Level name="Année" column="annee" uniqueMembers="true"/>
     </Hierarchy>
</Dimension>
```

Par exemple dans la dimension « Année », on supprime le niveau 'All' puisque les calculs des estimateurs sont non additifs sur cette dimension, mais on va préciser que le membre par défaut est non pas la première année 2005, mais l'avant dernière année stockée en base 2008 (nous verrons plus tard pourquoi cette année).
De la même façon l'attribut `hasAll` de la dimension « Stratification » prendra *`false.`*

```
<Dimension name="Composition du peuplement recensable G1">
    <Hierarchy hasAll="true" allMemberName="total" primaryKey="comp_r_id">
      <Table name="peup_comp_r"/>
      <Level name="resfeuil" column="resfeuil_r_libelle" uniqueMembers="true"/>
      <Level name="compsimp" column="compsimp_r_libelle" uniqueMembers="true"/>
      <Level name="comp_r" column="libelle" uniqueMembers="true"/>
    </Hierarchy>
    <Hierarchy name="Composition du peuplement recensable G2" hasAll="true"
               allMemberName="total" primaryKey="comp_r_id">
      <Table name="peup_comp_r"/>
      <Level name="compreg_r" column="compreg_r_libelle" uniqueMembers="true"/>
      <Level name="compsimp_r" column="compsimp_r_libelle" uniqueMembers="true"/>
      <Level name="comp_r" column="libelle" uniqueMembers="true"/>
    </Hierarchy>
    <Hierarchy name="Composition du peuplement recensable G3" hasAll="true"
               allMemberName="total" primaryKey="comp_r_id">
      <Table name="peup_comp_r"/>
      <Level name="comp_cg" column="comp_cg_libelle" uniqueMembers="true"/>
      <Level name="compsimp_r" column="compsimp_r_libelle" uniqueMembers="true"/>
      <Level name="comp_r" column="libelle" uniqueMembers="true"/>
    </Hierarchy>
</Dimension>
```

Une dimension peut contenir plus d'une hiérarchie. La dimension « Composition du peuplement recensable G1 » ci-dessous contient trois hiérarchies. La première met l'accent sur la séparation des feuillus/résineux, contrairement à la deuxième qui met l'accent sur la mixité, la troisième est un autre regroupement de composition mixte. Par défaut le nom de la première hiérarchie prend le nom de la dimension. Nous aurions pu créer des dimensions différentes, cependant nous ne voulons pas croiser dans les analyses ses trois grains. En les combinant dans la même dimension mais sur trois hiérarchies distinctes, le 'All' correspondant à la valeur 'Toutes les combinaisons', nous empêchons les applications de les utiliser simultanément dans la même requête.

3.2 Des schémas en étoile à la constellation

Jusqu'ici nous avons vu comment créer un cube à partir de la table de faits et des dimensions jointes à la table de faits. Cependant Mondrian nous offre d'autres possibilités afin d'améliorer notre schéma.

3.2.1 Dimensions partagées

Nous avons vu au paragraphe précédent, que les dimensions pouvaient être partagées par les cubes. Cela garantit une similarité dans la sortie des résultats. Les dimensions communes n'appartiennent pas à un cube, elles sont déclarées pour un schéma, il faut donc spécifier leur source de données (clé étrangère) au cube.

On définit une première fois dans l'en-tête les dimensions partagées :

```
<Dimension name="Altitude">
    <Hierarchy hasAll="true" allMemberName="total" primaryKey="clz">
        <Table name="clz"/>
        <Level name="Classe d'altitude" column="clz_libelle" uniqueMembers="true"/>
    </Hierarchy>
</Dimension>
```

Et dans chaque cube il suffit d'ajouter :

```
<Cube name="Point" defaultMeasure="nb_pointsP">
    <Table name="cube_point" />
    ...
    <DimensionUsage name="Altitude" source="Altitude" foreignKey="clz" />
</Cube>

<Cube name="Peuplier" defaultMeasure="Nombre de points Peuplier">
    <Table name="cube_peuplier" />
    ...
    <DimensionUsage name="Altitude" source="Altitude" foreignKey="clz" />
</Cube>
```

Cet exemple montre la dimension «Altitude » d'être joint :

- au cube « Point » en utilisant la clé étrangère cube_point.clz
- au cube « Peuplier » en utilisant la clé cube_peuplier.clz étrangers.

3.2.2 Cubes Virtuels

Nous voici donc avec sept cubes modélisés. La première démarche de l'utilisateur est donc de choisir dans quel domaine d'étude il désire faire son analyse, s'il veut forer ses requêtes dans les cubes « Point », « Forêt », « Arbre peuplier » ou « Landes ». Mais à ce niveau, il ne peut faire aucune comparaison entre ces objets d'inventaires. Cependant comme vu plus haut les protocoles forêt et peupleraie sont très proches. Une analyse comparative du type volume en bois énergie ou volume en bois d'industrie ne serait pas facile à réaliser avec cette modélisation (les requêtes devront se formuler dans chacun des cubes).

Les cubes virtuels peuvent résoudre ce problème. Un cube virtuel est une jointure entre deux cubes différents. Il est constitué en sélectionnant les mesures et les dimensions communes des cubes sous-jacents, les cubes joins ne doivent pas avoir le même nombre de dimensions. Les utilisateurs finaux voient le cube virtuel comme un nouveau cube. Parce qu'on ne stocke que leurs définitions et non pas leurs données, les cubes virtuels ne nécessitent dont pas un gros stockage physique mais plutôt de la mémoire. On peut donc utiliser les cubes virtuels pour créer des combinaisons et des variantes de cubes existants sans beaucoup d'espace supplémentaire.

Actuellement on a un cube par type élémentaire. Pour palier à ce problème, il reste à construire deux cubes virtuels « Forêt et Peupleraie » qui regrouperait les dimensions communes aux cubes « Forêt » et « Peupleraie » et « Arbre forêt et Peuplier » qui regrouperait les cubes « Arbre » et « Peuplier ». Les données communes des regroupements de ces types élémentaires seront chargées dans des cubes virtuels. La figure 4-15 ci-dessous résume cette association.

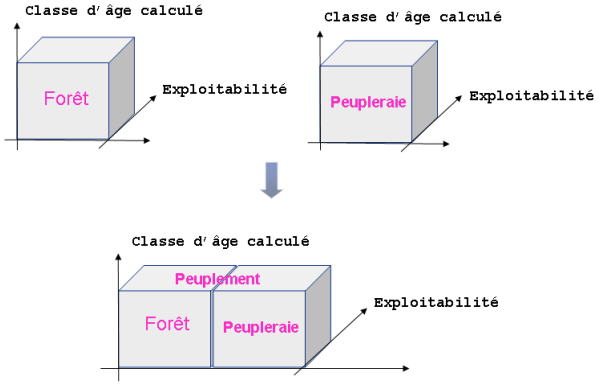


Figure 4-15. Cubes Virtuels « Forêt » et « Peupleraie »

Les dimensions « Classe d'âge calculé » et « Exploitabilité » sont communes aux deux cubes « Forêt » et « Peupleraie ». En joignant ces dimensions, on implémente un nouveau cube qui permettra d'analyser des informations pour ces deux types élémentaires.

```
<VirtualCube name="Peuplier et Foret" defaultMeasure="Nombre de points foret+peuplier">
  <VirtualCubeDimension name="Localisation"/>
  <VirtualCubeDimension name="Propriété"/>
  <VirtualCubeDimension name="Couverture du sol"/>
  <VirtualCubeDimension name="Altitude"/>
  <VirtualCubeDimension name="Croisement COUVERTURE x UTILISATION x TAILLE MASSIF"/>
  <VirtualCubeDimension name="Age calculé" />
  <VirtualCubeDimension name="Composition du peuplement recensable" />
  <VirtualCubeDimension name="Essence principale recensable" />
  <VirtualCubeDimension name="Exploitabilité" />
  <VirtualCubeDimension name="Mode de calcul de l'âge" />
  <VirtualCubeDimension name="Indicateur de peuplement recensable" />
  <VirtualCubeDimension name="Structure forestière" />

  <VirtualCubeMeasure cubeName="Peuplier" name="[Measures].[Surface Peuplier]"/>
  <VirtualCubeMeasure cubeName="Foret" name="[Measures].[Surface Foret]"/>
  <CalculatedMember name=" Foret + Peuplier " dimension="Measures">
    <Formula>[Measures].[ Surface Peuplier] + [Measures].[ Surface Foret]</Formula>
  </CalculatedMember>
</VirtualCube>
```

Le chargement du cube virtuel entraine le chargement des cubes sous-jacents.

Les cubes virtuels ont un autre atout, ils sont une solution pour deux problèmes. Un dès avantage du cube virtuel, est qu'il peut aussi se fondre sur un seul cube et ainsi exposer seulement des sous-ensembles sélectionnés de ses mesures et de ses dimensions. Cette démarche peut s'avérer utile afin de limiter l'accès de certains utilisateurs lors de la visualisation des cubes sous-jacents. Par exemple si certaines des informations d'un cube sont sensibles et ne conviennent pas à tous les utilisateurs, on peut créer un cube virtuel à partir du cube existant et supprimer les informations à cacher. Ensuite, il suffit de créer deux rôles : le premier contenant les utilisateurs autorisés à consulter ces informations sécurisées et le deuxième contenant les autres utilisateurs. Enfin, on accorde l'accès au cube de base aux utilisateurs « administrateur » et l'accès au deuxième cube aux autres. Cette solution est primordiale pour l'IFN. Beaucoup de données ont une utilité unique pour des experts techniques internes ou des données répondent à des demandes spéciales du ministère et nécessitent une certaine confidentialité.

3.3 Calculs des estimateurs

La dernière étape à la modélisation du schéma « dwifn » consiste à ajouter les estimateurs que les analystes de l'IFN recherchent. Pour cela nous allons utiliser les membres calculés. Ils fournissent des outils afin de définir une logique métier complexe : par exemple créer des mesures dont la valeur ne provient pas d'une ou plusieurs colonnes de la table de faits mais d'une formule MDX.

Remarquons tout d'abord comment MDX peut nommer au moins de trois manières différentes les cellules du cube de la figure 4-16 :

```
 (Propriété.Communale,Measures.Surface,Année.2007)
=(Propriété.Communale,Measures.Surface,Année.2008.PrevMember)
=(Propriété.Communale,Measures.Surface,Année.2006.NexMember)
```

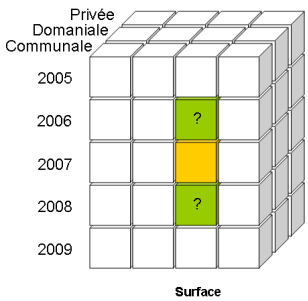


Figure 4-16. MDX - Nom des cellules

Les outils sont maintenant en place afin de calculer les estimations des cubes.

3.3.1 Calcul de la surface

Nous avons vu comment calculer la surface pour une année et une stratification (somme sur toutes les dimensions hormis la dimension « Année » et la dimension « Stratification » qui sont non-additives). Mais cette mesure ne correspond pas aux résultats voulus. Nous devons encore effectuer une moyenne temporelle. Par exemple nous cherchons à établir la surface calculée sur la période [2005, 2007]. Pour cela, nous la calculerons à partir de l'année médiane 2006. Nous connaissons la longueur de cette période qui est de 3 ans.

Nous avons recours à deux mesures calculées intermédiaires avant d'obtenir au final la surface 2006 moyenne annuelle sur 3 années.

La première mesure calculée nous permet de définir cette moyenne pour chacun des membres de la dimension « Année » :

```
<CalculatedMember name="tteSurf (A,A-1,A+1)" dimension="Measures" visible="false">
    <Formula>
( ([Année].currentmember,[Measures].[Surf]) +
  ([Année].currentmember.prevmember,[Measures].[Surf]) +
  ([Année].currentmember.nextmember,[Measures].[Surf])  ) / 3
    </Formula>
</CalculatedMember>
```

La surface 2006 est la somme des surfaces 2005, 2006 et 2007 divisée par 3.

Cette mesure répond presque à nos besoins. Cependant que se passe-t-il pour les années limites 2005 et 2009 ?

[Croisement COUVERTURE x UTILISATION x TAILLE MASSIF].[total].[FORET DE PRODUCTION]	[Measures].[Surf]	[Measures].[tteSurf (A,A-1,A+1)]	[Measures].[Surf (A,A-1,A+1)]
[Année].[2005]	15,257,800.717	10,073,587.1	
[Année].[2006]	14,962,960.584	15,079,770.618	15,079,770.618
[Année].[2007]	15,018,550.553	15,068,186.984	15,068,186.984
[Année].[2008]	15,223,049.816	15,176,661.846	15,176,661.846
[Année].[2009]	15,288,385.17	10,170,478.329	

Tableau 4-6. MDX - Résultats de la mesure calculée « tteSurf (A,A-1,A+1) »

Pour l'année 2005, la moyenne est calculée à partir des surfaces 2006, 2005 et 2004. Mais pour l'année 2004, la cellule de la surface est vide dans le cube, cela retourne une valeur nulle et par conséquent la moyenne est erronée.

Il faut donc pour les calculs des surfaces estimées à partir de 3 campagnes supprimer les années limites 2005 et 2009. Ce qui donne une nouvelle mesure calculée :

```
<CalculatedMember name="neutreSurf (A,A-1,A+1)" dimension="Measures" visible="false">
    <Formula>
    iif([Année].currentmember is [Année].[2005] or
            [Année].currentmember is [Année].[2009],
        null,
        [Measures].[tteSurf (A,A-1,A+1)])</Formula>
</CalculatedMember>
```

Si nous voulions estimer les variables sur 5 campagnes, il faudrait alors exclure les deux années limites 2005, 2006 et 2008, 2009.

Afin d'indiquer la fiabilité des résultats de l'estimateur de la surface, MDX permet aussi à l'aide d'une mesure calculée d'évaluer à l'aide d'une expression conditionnelle la valeur obtenue et d'afficher une mise en forme particulière suivant les résultats des tests.

```
<CalculatedMember name="nbptt" dimension="Measures" visible="true">
    <Formula>
        ( ([Année].currentmember,[Measures].[nb_pointsP]) +
          ([Année].currentmember.prevmember,[Measures].[nb_pointsP]) +
          ([Année].currentmember.nextmember,[Measures].[nb_pointsP])  )
    </Formula>
</CalculatedMember>
<CalculatedMember name="Surf (A,A-1,A+1)" dimension="Measures">
    <Formula>[Measures].[neutreSurf (A,A-1,A+1)]</Formula>
        <CalculatedMemberProperty name="FORMAT_STRING"
              expression="IIf(([Measures].[nbptt] > 50.0),
                              '|#,##0|style=green',
                             IIf(([Measures].[nbptt] > 10.0),
                                       '|#,##0|style=yellow',
                                       '|#,##0|style=red'))" />
</CalculatedMember>
```

	Mesures									
	Surf (A-1,A,A+1) Ha					nbptt (A-1,A,A+1)				
	Année					Année				
Localisation Administrative	2005	2006	2007	2008	2009	2005	2006	2007	2008	2009
-RHONE-ALPES		7 772	9 004	9 224		7	21	34	41	27
AIN		3 215	3 531	4 030		2	10	17	20	12
ARDECHE		698				1	1			
DROME		165	342	1 112			1	2	3	2
HAUTE-SAVOIE										
ISERE		2 934	3 115	2 466		3	8	9	11	6
LOIRE										
RHONE			123	322				1	2	2
SAVOIE		760	1 894	1 293		1	1	5	5	5

Sous-domaine : [us_nm=PEUPLERAIE]

Figure 4-17. Résultats dans JPivot de la surface des peupleraies en Rhône-Alpes

Par exemple, si le nombre de points obtenus est compris entre 10 et 50 points la cellule s'affiche en orange (l'échantillon n'est pas assez important) et si le nombre de points est inférieur à 10 alors la cellule s'affiche en rouge.

3.3.2 Calcul des autres variables quantitatives

Les calculs des estimateurs des autres variables quantitatives suivent le même principe. Par exemple calculons le volume sur pied sur 3 campagnes annuelles consécutives :

```
<Measure name="tfk_v" visible="false" aggregator="sum">
    <MeasureExpression>
        <SQL dialect="generic">
            ( cube_parbre.sk * cube_parbre.pfk * cube_parbre.mfk_v ) / 1E4
        </SQL>
    </MeasureExpression>
</Measure>
<CalculatedMember name="ttV (A,A-1,A+1)" dimension="Measures" visible="false">
     <Formula>( ([Année].currentmember,[Measures].[tfk_v]) +
                ([Année].currentmember.prevmember,[Measures].[tfk_v]) +
                ([Année].currentmember.nextmember,[Measures].[tfk_v])  ) / 3</Formula>
</CalculatedMember>
<CalculatedMember name="V (A,A-1,A+1)" dimension="Measures" visible="true">
     <Formula>
   iif([Année].currentmember is [Année].[2005] or [Année].currentmember is [Année].[2007],
       null,
       [Measures].[ttV (A,A-1,A+1)])</Formula>
  </CalculatedMember>
```

Ce qui donne avec l'estimation de la biomasse aérienne dans le cube « Arbre Peuplier » :

	Mesures	
	V (A,A-1,A+1) m³	BIOM_AR (A,A-1,A+1) tms
	Année	Année
Diamètre	2006	2006
-total	24 169 582,223	14 707 927,751
+GROS BOIS - D >= 37,5	12 084 648,978	6 977 931,139
+MOYEN BOIS - 22,5 <= D < 37,5	9 018 527,278	5 344 226,184
+PETIT BOIS - 7,5 <= D < 22,5	3 066 405,967	2 385 770,428

Figure 4-18. Résultats dans JPivot du volume et de la biomasse aérienne en France des peupliers

4 Résultats et Evolutivités

La mise en place du modèle physique a permis de concevoir le modèle logique implémenté sur le serveur décisionnel Mondrian. Il nous reste maintenant à voir comment exploiter la base de données multidimensionnelle, analyser ces avantages et vérifier si les objectifs demandés sont atteints.

4.1 Résultats dans JPivot

Le dernier niveau d'un système d'entrepôt de données est le client OLAP qui offre à l'utilisateur une interface avec des outils d'analyse interactive et de reporting. Son rôle est de découvrir des connaissances grâce à la seule visualisation et interaction avec les données. Le client OLAP le plus adopté est la table de pivot.

JPivot est un outil qui permet aux utilisateurs d'exécuter des navigations OLAP. Livré avec Mondrian, JPivot est une bibliothèque de balises JSP.

Son interface de navigation permet une exploration très interactive comparée à ce qui existe actuellement à l'IFN. La facilité de navigation avec la barre d'outils permet rapidement d'exécuter :

- Un drill-down afin d'afficher pour un niveau d'une hiérarchie d'une dimension donnée, tous les membres enfant sans afficher les membres parent : on descend dans la hiérarchie.
- Un slice afin d'afficher les données qui correspondent à un critère de « filtrage ».
- Un drill-through afin d'afficher en un clic tous les faits qui ont contribués à un croisement sélectionné.

L'interface de JPivot est composée d'une barre d'outils (figure 4-20) et d'une table. L'ensemble des boutons de la barre d'outils permet de manipuler les données. Ils permettent de gérer les dimensions du cube, l'affichage de la table JPivot, le mode de navigation, les graphiques et l'impression/export.

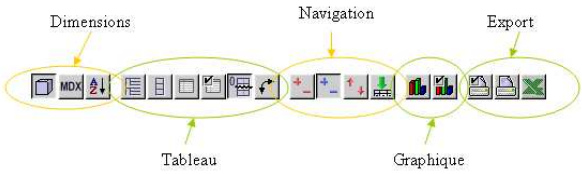


Figure 4-19. JPivot - Barre de Menu de JPivot

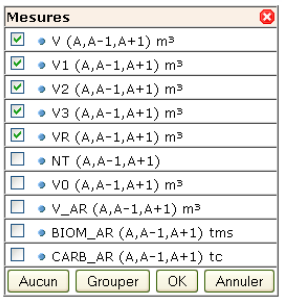


Figure 4-21. JPivot - Choix des mesures

Figure 4-20. JPivot - Choix des dimensions

Le premier bouton est le plus important : il permet l'exploration du cube en ajoutant/enlevant des champs ou en filtrant. Le panel de la figure 4-21 est divisé en trois sections permettant la configuration d'affichage des dimensions en colonnes, en lignes et celles qui peuvent éventuellement servir comme filtre. Un atout supplémentaire est le choix de l'ordre des dimensions en utilisant les flèches. Il est possible aussi de filtre les valeurs qui vont être affichées. Les valeurs peuvent elles aussi être triées.

Les autres boutons permettent des changements de présentation, fusion ou non des lignes et des colonnes, supprimer les lignes dont la valeur est nulle ou les conserver toutes, inter-changer les colonnes et lignes.

Un nouvel avantage dans le système d'entrepôt de données est le rôle des dimensions « Année » et « Stratification ». Actuellement la comparaison n'est pas possible entre les différentes années et entre les stratifications. Avec des outils OLAP la gestion des évolutions des surfaces et des volumes sur différents territoires devient plus facile. Il faut cependant faire attention dans l'exploitation de ce tableau, les résultats sont des estimations et l'intervalle de confiance n'est pas encore fourni. Les écarts entre deux estimations sont en général inférieurs à l'intervalle de confiance et ne permet donc pas de considérer l'évolution comme significative. Les seules évolutions qui peuvent être interprétées sont celles où les intervalles de confiance ne se recoupent pas, ce que le modèle ne permet pas de dire. Avec la mesure du comptage des points, l'outil proposé permet seulement de donner des indications concernant les évolutions mais aucune conclusion ne doit être tirée.

			Mesures									
			nb_pointsP					Surf(A) Ha				
			Stratification					Stratification				
			Stratification Ph2 par DEP, types FAO, PROpriete et Ph1					Stratification Ph2 par DEP, types FAO, PROpriete et Ph1				
Localisation Administrative			Année					Année				
(All)	Région	Département	2005	2006	2007	2008	2009	2005	2006	2007	2008	2009
France entière			6 829	6 533	6 878	6 238	6 122	15 257 800,717	14 962 960,584	15 018 550,553	15 223 049,816	15 288 385,17
France entière	CENTRE		460	457	470	418	407	896 068,218	916 002,32	909 849,567	909 399,496	920 149,225
	CENTRE	CHER	86	101	91	85	72	171 098,752	182 140,53	174 243,976	184 915,751	178 894,139
		EURE-ET-LOIR	37	31	40	39	31	64 870,798	69 280,956	73 893,043	85 130,211	69 043,604
		INDRE	63	47	61	61	50	119 567,179	106 654,966	131 351,568	130 478,899	115 033,501
		INDRE-ET-LOIRE	77	76	74	66	66	154 635,149	141 906,542	142 652,293	133 574,255	164 516,665
		LOIR-ET-CHER	98	114	111	91	102	203 841,907	217 448,201	208 468,461	211 501,753	209 303,831
		LOIRET	99	88	93	76	86	182 054,432	198 571,125	179 240,226	163 798,628	183 357,484

Sous-domaine : [us_nm=FORET DE PRODUCTION]

Figure 4-22. Ventilation par Année

La figure 4-22 montre un exemple d'une possible configuration de la table de pivot afin d'étudier l'évolution de la surface au cours du temps. La surface représentée dans cette grille est l'estimation annuelle de la surface en région Centre, toutes les années peuvent être prises en compte.

4.2 Bilan

Cette première partie de mise en place d'un entrepôt de données à l'Inventaire forestier national se résume ainsi : modélisation physique de la base qui va stocker et structurer les données, conception des cubes, mise en place des calculs afin de produire des résultats et enfin implémentation d'un client OLAP offrant aux utilisateurs une exploration conviviale et rapide. Cette action fut menée sur 4 mois, de janvier 2009 à avril 2009 et a abouti à une présentation à l'ensemble des directions de l'établissement. Elle a démontré avec succès l'accès à faire des recherches de manière très interactive, le besoin interne aux utilisateurs de forer rapidement sur les données, la possibilité d'analyser sur un axe temps. Cette phase va d'ailleurs se transformer en projet interne et rentrer en production.

Un dès premier objectif d'une structure décisionnelle implémentée à l'IFN est la production de résultats. Cet objectif est atteint. Les résultats fournis sont identiques à ceux issus de la chaîne de traitement, l'ensemble des variables y sont calculées. La structure de l'entrepôt de données peut-être considérée comme une base de production robuste et fiable.

Cependant avant la mise en production de cette nouvelle architecture et la diffusion des résultats en interne ou en externe, il reste à implémenter les tests de validité des résultats produits. Des tests de comparaison ont été faits visuellement : les mêmes requêtes sont exécutées dans les deux environnements Service de calcul et Mondrian et les lignes sont comparées via les interfaces d'OCRE et de JPivot. Les données en base ont aussi été contrôlées via des requêtes SQL en accès directs aux bases de données. Dans un souci d'une validation, il faut encore rendre ces tâches automatiques, généraliser les comparaisons à l'état de référence, opérer à des tests d'interface, mettre en place des requêtes afin de tester des cohérences existantes, etc.

Faute de temps les performances n'ont pas été testées, ce qui est un peu regrettable. Malgré tout comparer les résultats produits avec le service de calcul existant n'aurait pas été une tâche équitable. En effet celui-ci calcule en plus des estimations les coefficients de variation de ceux-ci (équivalent d'une variance), ce qui alourdit les temps de calcul. A cela, des simulations afin d'accroître grandement les tables de fait avec une dizaine de campagnes auraient permis de valider l'intérêt d'une mise en place des entrepôts de données à l'IFN.

A priori, la performance du couple Mondrian/JPivot semble correcte. Au cours des démonstrations réalisées, le serveur a tenu ses promesses en déployant rapidement des tableaux croisés dans chacun des cubes. En ce qui concerne JPivot, l'interface, comme toute interface web, perd en performance dès lors que le volume des données est trop grand, et ce n'est non pas pour des questions de performance du serveur de Mondrian.

Après cette première itération, quelques pistes d'évolution permettraient d'optimiser et de rendre encore plus robuste cette première version du système d'entrepôt de données. Remarqué dans la constitution des dimensions thématiques, une dès première action à réaliser est la diminution du nombre de dimensions. En effet les tables de fait en comportent trop, et leur volumétrie croît inutilement. C'est d'autant plus envisageable que beaucoup de données sont corrélées, d'autres servent d'indicateurs (et donc ont au plus 2 valeurs). Une autre piste consisterait à partager certaines dimensions : la dimension « Essence » et « Essence principale » ont exactement la même structure et les mêmes valeurs. Au lieu de créer deux tables de dimensions, une seule est implémentée.

Une autre piste d'optimisation, concerne les calculs. La dimension « Année » joue un rôle capital dans toutes les analyses. Mondrian permet de créer des tables d'agrégats et ainsi stockés des résultats agrégés intermédiaires. Nous pourrions construire une table d'agrégat par campagne d'inventaire. Cette solution reste toutefois à tester.

Démarche d'implémentation d'une Interface Web Olap Cartographique

Dans le précédent chapitre nous avons vu ce qu'offrait déjà l'interface de JPivot afin de piocher les résultats dans l'entrepôt de données de manière exploratoire, intuitive et interactive. Au chapitre 3, nous avons décrit les motivations des utilisateurs afin de disposer d'applications géodécisonnelles qui permettraient de visualiser spatialement les résultats produits. Aussi pour répondre à ces nouveaux besoins, la présente partie du rapport aborde le développement d'une interface permettant d'exploiter les dimensions géographiques des cubes, de déployer des nouvelles fonctionnalités SIG et de synchroniser l'analyse des données calculées représentées parallèlement sous forme de carte et sous forme de tableau. Dans un premier temps, nous décrirons l'architecture choisie et mise en place de notre prototype, comment elle s'appuie sur l'entrepôt de données et les cubes décrits précédemment. Puis, nous présenterons notre réalisation en détail et le fonctionnement de cette nouvelle approche. Enfin, nous conclurons ce chapitre par un bilan des caractéristiques obtenues et les limites, en rapport avec les objectifs visés.

1 Architecture de l'application

Nous commençons par décrire l'architecture gérérale de la maquette qui a été implémentée, tout d'abord d'un point de vue globale avec les connections entre les différents composants, puis plus précisément nous exposerons les caractéristiques et les fonctionnalités de chaque niveau.

1.1 Architecture générale

Comme nous l'avons déjà exprimé précédemment, l'IFN a déjà développé et mis en place un module permettant de créer des applications produisant simultanément l'affichage de cartes et de tableaux. Ce module produit des résultats de données brutes ponctuelles, c'est-à-dire que les résultats sont issus de requête type sélection dans une base de données relationnelle qui contient les informations collectées. Ces données sont attachées au point d'une grille systématique et aucune estimation n'est calculée en ligne. Ce module est implémenté sous forme d'une bibliothèque applicative générique et a fait ses preuves en étant implanté dans déjà trois applications en production (EFOREST[30], RTM[31] et BDN[32]). Son architecture est donc fiable. Les points communs de ce module avec notre prototype concernent la base de données avec sa cartouche spatiale, un serveur web simple muni de l'extension PHP et d'un serveur cartographique afin de générer les cartes. Il est donc logique que l'architecture que nous allons établir ressemble à celle-ci.

La solution web réalisée repose sur une architecture 3-tiers, comme le montre la figure 5-1 ci-dessous. Elle est composée :

- d'une base de données spatiale dans laquelle sont stockées les données spatiales et alphanumériques PostgreSQL muni de la cartouche spatiale Postgis
- de deux serveurs :
  - le serveur ROLAP Mondrian traite et gère les requêtes multidimensionnelles
  - le serveur cartographique MapServer permet d'afficher les données géographiques
- d'un client web dans lequel l'utilisateur peut visualiser les résultats tabulaires et cartographiques. Ce dernier intègre un client SIG GeoExt et un client OLAP JPivot.

[30] EFOREST : Réalisation d'une plate forme d'import de données forêt au niveau Européen.
[31] RTM : Risques en Terrain de montagne, gère de façon centralisée les risques en terrain de montagne.
[32] BDN : Application qui fournit aux agents de l'ONF une gestion centralisée des données naturalistes.

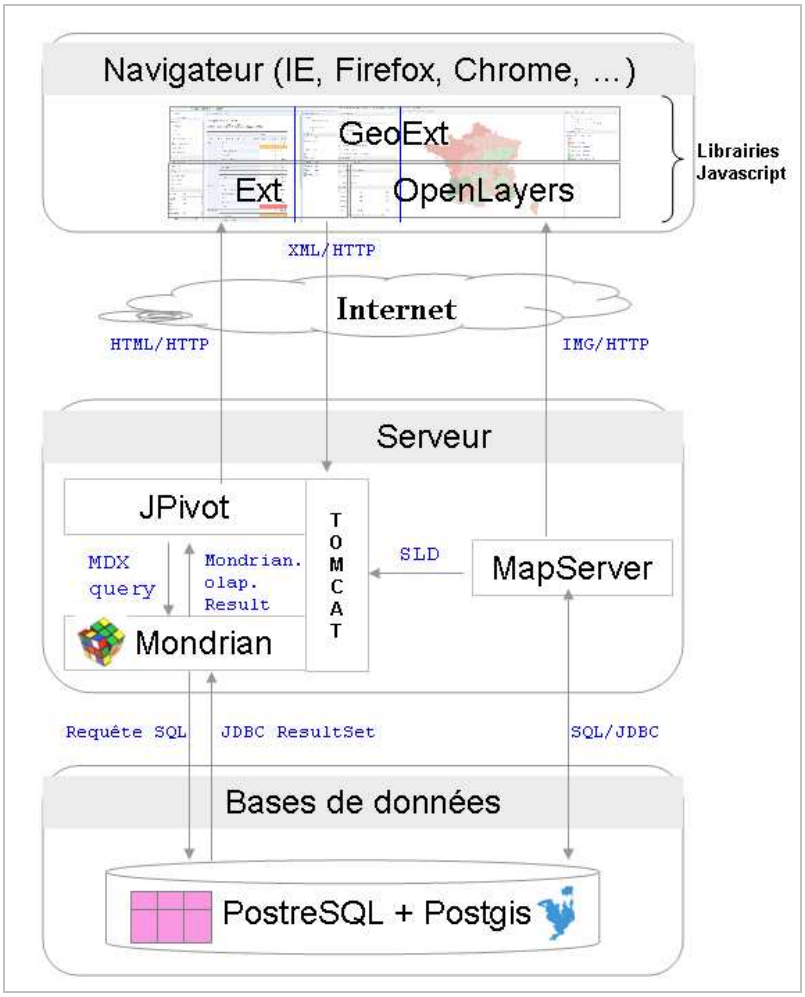


Figure 5-1. Architecture de la solution

Mondrian ne permet pas de retourner une couche géographique, et donc de s'interfacer avec MapServer. Par la connaissance des métadonnées et du résultat, le serveur ROLAP peut extraire les informations utiles (identifiant de la session utilisateur (un layer par session), la liste des contours avec le résultat et la couleur associée par discrétisation), les enregistrer dans une table utilisateur et ainsi permettre au serveur cartographique d'extraire le layer de la carte de résultat.

1.2 Interface utilisateur

Cette architecture permet donc d'afficher simultanément la carte et le tableau résultant de la requête multidimensionnelle de l'utilisateur.

Grâce aux fonctionnalités d'Ajax, l'utilisateur travaille sur une unique page. Les différentes vues sont rafraîchies suivant les actions demandées. La figure 5-2 ci-dessous décrit succinctement ces interdépendances.

Figure 5-2. Composants de l'interface cliente

Depuis son poste client, l'utilisateur valide son choix du cube (1) qui s'affiche dans (2) et qui permet d'initialiser le formulaire (3). Ainsi il peut initialiser les membres des dimensions alphanumériques, de la dimension spatiale et les mesures qu'il désire paramétrer. La requête est envoyée au serveur ROLAP. Celui-ci l'exécute via la requête MDX élaborée à partir des choix de l'utilisateur. Le résultat est transféré dans les différents onglets du TabPanel (4).
En particulier, le résultat dans le TabPanel est :

- enregistré dans une table, ce qui permet au serveur cartographique d'afficher le layer de la carte résultat (premier onglet)
- affiché en tableau par le client JPivot (second onglet)
- envoyé à la vue « MDX » (troisième onglet)

La navigation dans la dimension spatiale interagit avec la carte via le formulaire. Les opérateurs multidimensionnels et les métadonnées de l'entrepôt de données sont accessibles via celui-ci de manière interactive. Les fonctionnalités SIG présentent ne sont pas très nombreuses : zoom de l'échelle et déplacement de la fenêtre et elles ne déclenchent pas d'opérations de navigation multidimensionnelles.

1.3 Entrepôt de données spatiales

L'entrepôt de données (nommé dwifn), définit dans le précédent chapitre, est implémenté entièrement dans le système de gestion de base de données relationnel PostgreSQL avec sa cartouche spatiale Postgis. Ce serveur présente l'avantage de s'installer sur tous les systèmes d'exploitation Unix, Linux ou Windows, de façon très conviviale. L'administration et la configuration de la base se font par l'interface très pratique PgAdmin III.

De plus le serveur est étendu avec des fonctionnalités spatiales grâce à la cartouche Postgis. Celle-ci permet l'indexation des colonnes spatiales grâce aux index de type GIST (Generalized Search Tress). PostgreSQL supporte les index de type R-Tress, mais leur implémentation est moins puissante que celle de type GIST et ne doit donc pas être utilisé. L'index GIST offre deux avantages :

- il permet l'indexation de valeurs nulles (cet intérêt ne nous concerne pas, nous n'avons aucun membre dont le contour géographique est à nul)
- il permet d'indexer de gros objets en ne prenant en compte que leur bounding box, ce que ne ferait pas un index R-Tree, limitant l'indexation des objets de moins de 8Ko.

Les granularités les plus fines des dimensions géographiques ont peu de cardinalités (96 départements ou 309 régions forestières), les performances de l'application sont correctes, il n'a pas été nécessaire de travailler sur l'optimisation pour les données spatiales.

Plus précisément les données alphanumériques et spatiales sont stockées en utilisant le schéma en constellation décrit dans le chapitre précédent. Les contours géographiques des deux dimensions

spatiales (la dimension administrative et la dimension écologique) sont enregistrés comme attribut à partir de fichiers Shape via la fonction « shp2pgsql ».

L'entrepôt de données contient aussi un ensemble de métadonnées, surcharge de la base de données MetaIFN qui permet d'opérer le chargement de l'entrepôt à partir des bases de données d'exploitation, de stratification et des fichiers SIG. Ces métadonnées identifient aussi les dimensions géographiques, les dimensions de type placette (opposées à celles qui sont de type arbre), les données, les hiérarchies définies au cours de cette réalisation, etc.

Afin de restituer l'affichage de la carte, MapServer récupère la couche Postgis dans une table utilisateur (son script est décrit dans la figure 5-3).

```
CREATE TABLE spatialuser.resultlayer
(
  id serial NOT NULL,
  sessionid character varying(64) NOT NULL,
  nom_geom character varying(100),
  codecouleur character(3) DEFAULT '001'::bpchar,
  modalitesymbole character varying(100),
  resultat double precision,
  the_geom geometry,
  CONSTRAINT pk_resultlayer PRIMARY KEY (id),
  CONSTRAINT enforce_dims_the_geom CHECK (ndims(the_geom) = 2),
  CONSTRAINT enforce_geotype_the_geom CHECK (geometrytype(the_geom) = 'MULTIPOLYGON'::text OR the_geom IS NULL),
  CONSTRAINT enforce_srid_the_geom CHECK (srid(the_geom) = (-1))
)
WITH (
  OIDS=FALSE
);
ALTER TABLE spatialuser.resultlayer OWNER TO etoile;
```

Figure 5-3. Définition de la table USERGEOM

Après discrétisation des résultats de la requête MDX, les valeurs sont ajoutées pour chaque géométrie retournée dans la colonne RESULTAT. L'analyse supplémentaire par symbole n'a pas été implémentée faute de temps, mais la colonne prévue à cet effet MODALITESYMBOLE enregistrerait les modalités de cette ventilation. Le code couleur du fichier SLD est saisi dans CODECOULEUR, les géométries sont dans THE_GEOM, respectivement les noms des géométries sont dans NOM_GEOM. Toutes ces informations sont rattachées à la session de l'utilisateur.

Un processus automatique de nettoyage extérieur permet de vider régulièrement cette table qui sinon est vouée à croître indéfiniment.

1.4 Serveurs

Le second niveau comporte deux serveurs, le plus important le serveur Rolap Mondrian et le serveur cartographique MapServer présentés maintenant.

1.4.1 Serveur Rolap Mondrian

Le serveur Rolap utilisé est Mondrian. Comme nous l'avons vu dans le chapitre 2, ce serveur OLAP open source, développé en Java, s'appuie sur une technologie relationnelle. Il établit le lien entre le client AJAX et le SGBD en traduisant les requêtes multidimensionnelles MDX générées en requêtes SQL.

Le couplage de Mondrian avec JPivot s'installe la plupart du temps en même temps. La configuration est très facile. Les paramètres de connexion se font dans le fichier web.xml :

```
<context-param>
    <param-name>connectString</param-name>
    <param-value>
Provider=mondrian;Jdbc='jdbc:postgresql://localhost:5432/dwifn';Catalog=/WEB-
INF/queries/dw5campagne.xml;JdbcDrivers=org.postgresql.Driver;JdbcUser=xxx;JdbcPassword=xx
x;PoolNeeded=false;
</param-value>
</context-param>
```

Le serveur Mondrian a l'importante tâche de gérer les dimensions et les mesures des cubes. Ceux-ci sont référencés dans un catalogue. Les cubes déployés sont ceux du chapitre précédent mais allégés. Ils sont décrits dans le catalogue *dw5campagne.xml*.

Certaines dimensions n'auraient pas de sens dans ce genre d'interfaces. Nous avons donc exclues la dimension de stratification et la dimension année, et nous avons gardées uniquement celles qui sont valides sur 5 campagnes. Les résultats de cette application concernent uniquement les estimations calculées sur les 5 dernières campagnes d'inventaire. Afin de simplifier l'interface les mesures ont été renommées :

```
Volume Forêt 2006 (2005-2007) => Volume Forêt
Biomasse Aérienne Peuplier 2006 (2005-2007) => Biomasse Aérienne Peuplier
Nombre de point => n'existe plus
…
```

Certaines dimensions ont héritées de propriété afin de leur donner un rôle et des fonctionnalités :

- Les dimensions géographiques ont la propriété « géométrie » qui a pour valeur le contour géométrique du membre.

```
<Property name="geometrie" column="dep_geom" />
```

- Les dimensions pouvant servir comme autre axe d'analyse (restitution cartographique sous forme de symbole) à l'axe géographique hérite de la propriété « symbole ».

```
<Level name="Classe d'altitude" column="libelle" uniqueMembers="true">
    <Property name="symbole" column="libelle" />
</Level>

<Level name="Statut forestier" column="sns_libelle" uniqueMembers="true">
                    <Property name="symbole" column="sns_libelle" />
</Level>
<Level name="Propriété" column="libelle" uniqueMembers="true"/>
```

- Afin de restreindre les domaines d'études des cubes, certaines dimensions, (peu nombreuses pour des raisons d'esthétisme, le panel listant ces critères n'est pas grand), héritent de la propriété « filtre ».

```
<Hierarchy hasAll="true" allMemberName="total" primaryKey="comp_r_id">
      <Table name="peup_comp_r"/>
      <Level name="resfeuil" column="resfeuil_r_libelle" uniqueMembers="true">
        <Property name="filtre" column="resfeuil_r_libelle" />
      </Level>
      <Level name="compsimp" column="compsimp_r_libelle" uniqueMembers="true">
        <Property name="symbole" column="compsimp_r_libelle" />
      </Level>
      <Level name="comp_r" column="libelle" uniqueMembers="true">
        <Property name="symbole" column="libelle" />
      </Level>
</Hierarchy>
```

Olap4j est une API java qui permet de construire des applications OLAP (Olap4J est aux données multidimensionnelles ce que JDBC est aux données relationnelles). Une application OLAP interagit avec un serveur OLAP via des requêtes MDX à travers des connexions. L'énoncé de la requête est défini suivant des métadonnées que les applications ont besoin de manipuler. Cette bibliothèque le permet par ses multiples objets. Elle peut aussi parcourir les catalogues des cubes, gérer les drivers et manager le cycle de vie des connections et des requêtes. En Annexe 2, un exemple d'appel à cette API est présenté.

Dans l'architecture mise en place, Mondrian tient un rôle central. Ses principales actions sont résumées ci-dessous :

- il initialise les listes déroulantes du formulaire client. Comme nous venons de le voir l'API Olap4j permet de construire les différentes listes (liste des cubes, liste des dimensions géographiques avec hiérarchies et niveaux, liste des dimensions possibles pour une analyse par symbole, liste des dimensions pouvant restreindre le domaine d'étude (opérations slice et dice)) à partir du catalogue (dw5campagne.xml).

- il exécute la requête MDX construite à partir des paramètres retournés en POST. Par défaut, les résultats en base et les paramètres du formulaire correspondent à ceux de « France entière » et la valeur est à zéro. Lorsque l'utilisateur valide ses choix, la requête se construit de la sorte : le *From* est initialisé par le choix du cube, le niveau d'une hiérarchie géographique constituera la ventilation ligne, si un niveau supplémentaire pour une analyse par symbole est choisie, il constituera une deuxième ventilation (construction d'un tuple, suite de plusieurs membres avec un CROSSJOIN : entre parenthèse et séparé par une virgule; on ajoute un NON EMPTY afin de ne pas récupérer les lignes vides) de la ligne, la variable quantitative déterminera la ventilation colonne, et si les niveaux des filtres ont des valeurs ils définiront le where de la requête.

- il déploie le tableau JPivot. JPivot est une bibliothèque de balises JSP (WCF[33]) qui permet, avec une requête multidimensionnelle MDX d'afficher les résultats sous un tableau et peut ajouter une barre d'outil. Dans notre implémentation nous n'avons pas intégrer la barre d'outil, seul le tableau final est affiché dans un onglet d'un panel. Cette solution (pourquoi pas une évolution dans une prochaine version) aurait été possible mais elle aurait nécessité de récupérer les actions de la barre ainsi que les actions du tableau (par exemple les drill-down, roll-up).

- il initialise le layer résultat dans la table USERGEOM dans Postgis. Avec l'API Olap4J le résultat est ajouté en base après le processus de discrétisation. Cette opération est implémentée pour l'instant en dur dans le code, mais il est tout à fait envisageable que l'utilisateur puisse choisir son modèle de classification à la volée.

- il construit le fichier SLD. La conception des cartes choroplèthes se produit en deux temps. Tout d'abord chaque résultat après discrétisation est identifié en base suivant un code couleur (nous verrons dans le paragraphe suivant comment ce code est pris en compte). Puis il convient de réécrire le fichier SLD, non pas que le style soit modifié, mais les valeurs des libellés des intervalles le sont (par exemple le code « 001 » correspond à une ventilation par région forestière de la surface des forêts de production soit l'intervalle [0, 500 ha[, puis dans une seconde requête le code « 001 » décrit la première échelle des valeurs du volume boisée pour ces mêmes régions soit [0, 10 000 m³[).

Les deux dernières actions que nous venons de décrire sont dédiées au serveur cartographique, nous allons maintenant décrire comment elles sont implémentées.

[33] Web Component Framework

1.4.2 Serveur MapServer

Afin de déployer des cartes dynamiquement dans nos interfaces, nous avons utilisé le serveur cartographique MapServer (Cf. chapitre 2, paragraphe 2.2), Mondrian ne permettant pas d'en produire. MapServer est un environnement de développement libre permettant de construire des applications internet à référence spatiale. Ce n'est pas un SIG complet, mais son rôle est suffisant pour notre implémentation et très simple à déployer.

Le MapFile est le fichier applicatif de configuration (généralement avec une extension .map). Il inclut les données à utiliser dans notre application, mais aussi les informations sur la manière de dessiner la carte, la légende. Dans notre MapFile nous avons configuré sur ce serveur quatre couches valables pour tous les résultats produits via le serveur Mondrian. Ces layers sont tous construits à partir de la table USERGEOM définit dans Postgis. La requête fait un select sur la colonne (THE_GEOM) avec une projection dans un plan (« using SRID = » figure 5-4). Afin de récupérer uniquement les résultats de l'utilisateur, la requête filtre les lignes grâce à l'identifiant de la session de l'utilisateur passée en paramètre lors de l'appel au serveur (ligne « FILTER » figure 5-4).

- La couche des résultats : c'est la principale et est obligatoire. Le layer est construit à partir des résultats enregistrés dans la table utilisateur USERGEOM.

```
LAYER
   NAME "carte_resultat"
   STATUS on
   TYPE polygon
   CONNECTIONTYPE postgis
   CONNECTION "user=postgres dbname=dwifn password=xxx host=localhost"
   DATA "the_geom from ( select id as gid, the_geom, codecouleur, sessionid
                         from spatialuser.usergeom
                       ) as alias using unique alias.gid using SRID = -1 "
   FILTER "alias.sessionid = '%EX%'"
   PROJECTION
        "init=epsg:27572"
   END
END
```

Figure 5-4. Layer "Résultat"

Les trois autres layers sont plus d'ordre esthétique. Ils n'engendrent pas de modification au niveau du contenu de la carte.

- La couche « Libellé » permet d'afficher les libellés des modalités géographiques. Au niveau le plus fin des dimensions géographiques, cette information est très vite illisible. La requête ne change pas, il suffit de spécifier le tag LABELITEM (figure 5.5).

```
LAYER
    NAME "carte_libelle"
    CONNECTIONTYPE postgis
    CONNECTION "user=postgres dbname=dwifn password=xxx host=localhost"
    DATA "the_geom
          FROM spatialuser.usergeom
           using unique id using SRID=-1"
    FILTER "sessionid = '%EX%'"
    STATUS on
    TYPE ANNOTATION
          LABELITEM nom_geom
          CLASS
              LABEL
                    ANGLE auto
                    SIZE 8
                    COLOR 0 0 0
                    TYPE truetype
                    FONT arial
              END
    END
END
```

Figure 5-5. Layer "Libellé"

- La couche « Contour » permet de dessiner les contours des zones géographiques. La requête d'extraction change un tout petit peu : la géométrie retournée est : `Boundary(the_geom)` fonction Postgis qui retourne le contour fermé du graphe de l'objet géométrique.

```
LAYER
    NAME "carte_contour"
    STATUS on
    TYPE LINE
    CONNECTIONTYPE postgis
    CONNECTION "user=postgres dbname=dwifn password=xxx host=localhost"
    DATA "the_geom from ( select Boundary(the_geom) as the_geom,
                                 id, sessionid, nom_geom, resultat
                          from spatialuser.usergeom ) as xx
        USING UNIQUE id USING SRID=-1"
    FILTER "sessionid = '%EX%'"
    PROJECTION
        "init=epsg:27572"
    END
    CLASS
        OUTLINECOLOR 128 128 128
    END
END
```

Figure 5-6. Layer "Contour"

- La couche « Centroïde » affiche le barycentre de l'entité géométrique sous forme de point géométrique avec la fonction `centroid(the_geom)`. L'affichage en carré penché se configure au niveau du style en ajoutant un symbole dans le fichier de configuration d'extension .sym et appelé dans le mapfile avec SYMBOLSET.

```
LAYER
    NAME "carte_centroide"
    STATUS on
    TYPE POINT
    CONNECTIONTYPE postgis
    CONNECTION "user=postgres dbname=dwifn password=xxx host=localhost"
    DATA "the_geom from ( select centroid(the_geom) as the_geom,
                                 id, sessionid,
                                 nom_geom, resultat
                          from spatialuser.usergeom ) as xx
          USING UNIQUE id USING SRID=-1"
    FILTER "sessionid = '%EX%'"
    PROJECTION
        "init=epsg:27572"
    END
    CLASS
        NAME "centroides"
        STYLE
            SYMBOL "carre-penche"
            SIZE 8
            COLOR 255 255 0
            OFFSET 1 -7
        END
    END
END
```

Figure 5-7. Layer "Centroïde"

Il nous reste, dans cette présentation des différents niveaux, à voir comment le niveau client s'interface avec le niveau des serveurs.

1.5 Client OLAP et Client cartographique

Le client principal est écrit avec la bibliothèque javascript GeoExt, fusion des librairies OpenLayers et Ext (Cf. chapitre 2.3). Le client JPivot n'est quasiment plus pris en compte. Avec l'avènement des librairies Ajax, il était primordial d'apporter dans l'implémentation de notre maquette, des interfaces cartographiques riches. JPivot, bien qu'il apporte une grande interactivité dans l'interface, manque de souplesse et de légèreté. Par exemple en effectuant une opération de drill-down, la requête est à nouveau exécutée et la page est entièrement rafraîchie. Nous avons donc opté pour la librairie mixte GeoExt non encore implémenté à l'IFN.

L'objet au cœur de cette application est le MapPanel. Né de la fusion du composant Map d'OpenLayers et l'objet Panel d'Ext, c'est lui qui aura la charge de construire la carte ainsi que son conteneur (panel).

```
var mapGeo = new OpenLayers.Map('map', {
    controls: [],
    allOverlays: false,
    'resolutions' : [1763.88793639, 1058.33276183, 705.555174556, 352.777587278,
                     176.388793639, 88.1943968195, 35.2777587278, 17.6388793639,
                     8.81943968195, 3.52777587278],
    'projection'  : 'EPSG:27572',
    'units'       : 'm',
    'tileSize'    : new OpenLayers.Size(500,500),
    'maxExtent'   : new OpenLayers.Bounds(-82714, 1443381, 1328396, 2854491)
});

// Layer Résultat
var layerGraphRes = new OpenLayers.Layer.WMS("Carte",
    "http://127.0.0.1:8085/cgi-bin/mapserv.exe/?map=C:/DEV/MAPSERV/localisation.map", {
            layers: "carte_resultat",
            format: "png",
            EX: "<% out.print(session.getId().trim()); %>",
            sld : "http://127.0.0.1:8083/mondrian/sld/sld.xml",
     }, {
            buffer: 0,
            isBaseLayer: true,
});

mapPanel = new GeoExt.MapPanel({
    id: 'carte',
    title: 'Carte',
    // Open Layer Map contenant les 4 layers de l'application
    map: mapGeo,
});
```

Ce panel sera intégré plus tard dans un tabPanel (tableau avec onglet). Le prototypage de l'interface est très rapide avec cette librairie. On instancie d'abord un OpenLayers.Map, ce qui n'est pas obligatoire pour le MapPanel, mais c'est mieux pour le customiser. On crée alors les OpenLayers.Layer, plus particulièrement des WMS layers pour obtenir les images des cartes. On crée ensuite le MapPanel auquel on ajoute le map.

Deux autres composants ont été très utiles et ont facilités l'implémentation de l'interface. Ce sont le LegendPanel et le TreePanel. Ces derniers se définissent très rapidement dès que le MapPanel est conçu :

```
var legendPanel = new GeoExt.LegendPanel({
    title: 'L&eacute;gende',
    iconCls:'globe',
    bodyStyle: 'padding:5px 0px 0px 10px',
});

var layerTree = new Ext.tree.TreePanel({
    title: 'D&eacute;coupage',
    iconCls:'globe',
    root: layerList
});
```

Le formulaire est un formulaire standard d'Ext. Il utilise les composants classiques des formulaires : `Ext.form.RadioGroup`, `Ext.form.ComboBox`, et `Ext.Toolbar`.

2 Interface visuelle

Lorsque l'utilisateur se connecte à l'application, celle-ci se résume à une unique interface. Elle est composée de quatre panneaux : le Tab Panel au centre, le Menu Panel en haut, le FormPanel à gauche et le LegendePanel à droite (figure 5-8 ci-dessous).

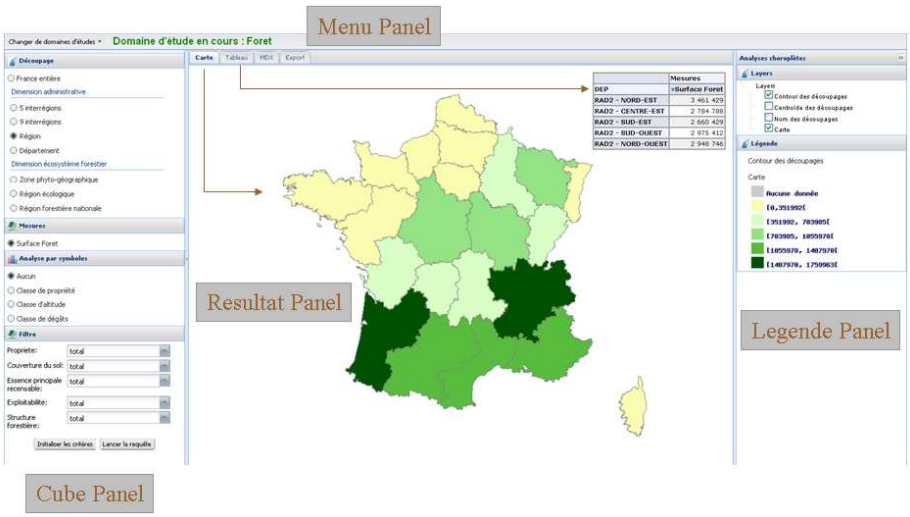


Figure 5-8. Interface visuelle

Le ResultatPanel, composé du MapPanel et du tableau JPivot, et le LegendePanel permettent de visualiser l'hypercube. Le premier onglet déploie le pouvoir expressif des cartes en mettant en valeur les dimensions géométriques, le second onglet affiche les tables multidimensionnelles de JPivot en utilisant surtout les autres dimensions. Le Menu Panel et le Cube Panel permettent de personnaliser les composants visuels du ResultatPanel.

2.1 Le Menu Panel

C'est la première étape par laquelle passe l'utilisateur : le choix du cube. L'étape n'est pas obligatoire, puisque par défaut un cube est initialisé par le plus commun celui des peuplements forestiers.

Cette démarche est différente de toutes les applications IFN qui produise des résultats statistiques. En effet le choix du domaine d'étude arrive le plus souvent après le choix du domaine spatial et celui des campagnes d'inventaire. L'approche dans les entrepôts de données en cela diverge.

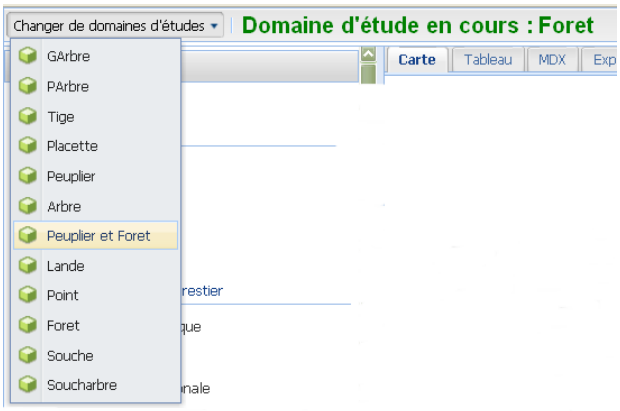


Figure 5-9. Menu ou Comment changer de cubes

Tous les panels dépendent de ce choix. Cette action, qui peut s'effectuer à tout moment, réinitialise le titre, le formulaire (liste des variables, liste des critères, …), la carte et le tableau multidimensionnel.

Afin de marquer ce changement, il était judicieux d'implémenter cette activité dans une barre de menu, elle équivaudrait à un changement de page dans une application web.

A la droite de la liste déroulante des cubes, l'intitulé du non du cube s'affiche en titre dans cette barre de menu.

2.2 Le Cube Panel

Le cube Panel permet de personnaliser les composants visuels du ResultatPanel et du LegendePanel. Il est construit comme un formulaire, ces champs implémentent les fonctionnalités de la barre à outil SOLAP.

Figure 5-10. Choix Géographique

La première rubrique du formulaire (figure 5-10) est disponible car le schéma multidimensionnel comporte deux dimensions spatiales.

L'interface ne permettant pas de croiser plusieurs couches cartographiques, et l'option pour une seule dimension spatiale avec des hiérarchies alternatives n'ayant pas été choisie, l'interface propose, par des boutons radio, un choix unique d'un niveau d'une des dimensions spatiales.

Ce sont les membres des niveaux des dimensions spatiales qui ont une propriété cartographique : leur contour géométrique. Les niveaux sont donc automatiquement identifiés par cette propriété qui porte toujours le même nom « géométrie ».

La seconde rubrique du formulaire (figure 5-11) présente l'ensemble des mesures disponibles pour le cube choisi.

Les dimensions « Année » et « Stratification », n'étant pas paramétrables et non additives, ces mesures calculent les estimateurs moyens sur les cinq dernières campagnes, l'année représentative est donc 2007 et la stratification par défaut est « Stratification Ph2 par DEP, types FAO, Propriété et Ph1 ».

Le choix de la mesure est unique.

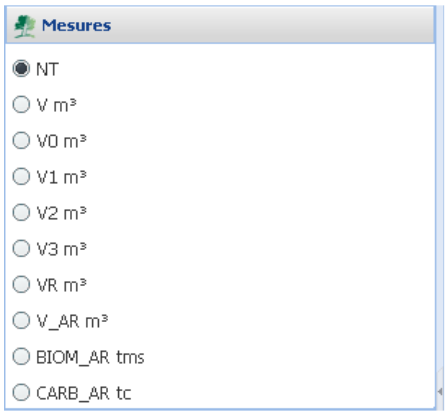


Figure 5-11. Choix de la mesure

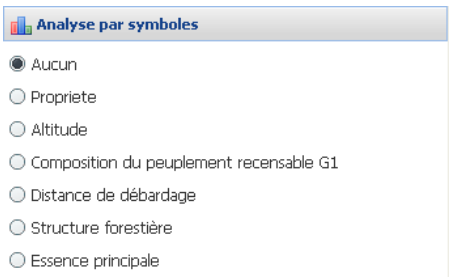


Figure 5-12. Choix d'une analyse supplémentaire par symbole

L'analyse supplémentaire par symbole n'a pas été intégrée faute de temps. Les techniques en amont sont cependant en place.

La troisième rubrique du formulaire (figure 5-12) permet de choisir, si l'utilisateur le désire, une ventilation supplémentaire dans son analyse.

La liste des critères proposés est très réduite pour des raisons d'affichage. Ce sont des dimensions qui ne possèdent qu'un faible nombre de modalités.

La requête MDX tient compte de cette dimension supplémentaire, les résultats sont enregistrés en base. La dernière phase qui restait à implémenter est l'affichage du layer graphique. MapServer inclut maintenant la possibilité de dessiner automatiquement des camemberts et des barres graphiques. Il suffit d'ajouter un layer supplémentaire, dont la source de données est la géométrie du résultat du membre géographique, avec une classe pour chaque part du camembert ou barre du graphique.

La quatrième rubrique du formulaire (figure 5-13) permet de restreindre les valeurs d'une ou plusieurs dimensions.

Afin d'alléger l'interface seules les dimensions les plus souvent utilisées ont été retenues.

Les listes déroulantes affichent les membres du niveau le plus en-dessous de la dimension. Ce composant du formulaire ne permet pas d'étendre la notion de hiérarchies. Une meilleure approche consisterait à afficher un popup avec la liste de l'ensemble des hiérarchies et des niveaux de la dimension de la même façon que JPivot. L'utilisateur filtrerait son analyse en sélectionnant ses membres à travers cette liste.

Par exemple si l'utilisateur désire faire une analyse dont la couverture au sol est la forêt fermée, il sélectionne cette valeur dans la liste. Pour l'instant le choix est unique, mais c'est envisageable de le rendre multiple.

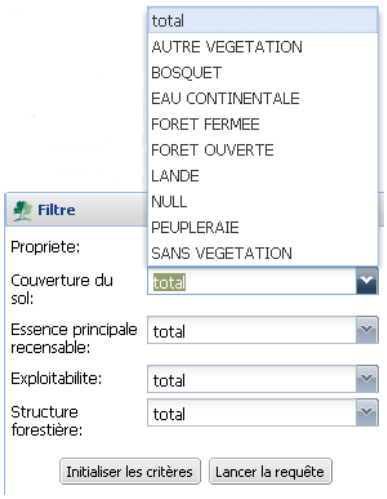


Figure 5-13. Choix Filtre

Le CubePanel est un formulaire classique. L'utilisateur doit à chaque fonctionnalité modifiée, valider son formulaire en cliquant sur le bouton « Lancer la requête ».

2.3 Le LegendePanel

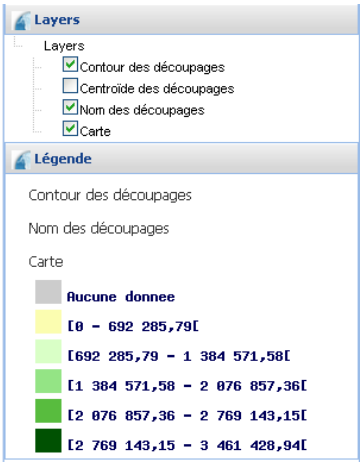


Figure 5-14. Legend Panel

Le LegendePanel est constitué de deux composants GeoExt.

Le LayerTree affiche la liste des couches cartographiques définis dans le mapfile et ajoutés dans le composant Map. L'utilisateur peut les restreindre. Par exemple le layer des libellés des membres du niveau spatial « département » est surchargé et donc rend la carte illisible.

Le LegendePanel présente les couches cartographiques sélectionnées ainsi que les couleurs et les seuils de chacune des classes issues de l'opération de discrétisation.

L'opération de discrétisation réalisée consiste à subdiviser équitablement l'intervalle de 0 à la plus grande valeur retournée en 5 classes d'étendues égales.

2.4 Le Resultat Panel

Le ResultatPanel est en réalité un TabPanel comportant quatre onglets. Il combine les différentes formes de résultat : l'affichage cartographique, l'affichage tabulaire et les affichages

graphiques (non implémenté), la quatrième page sert plus au débogage et affiche la requête MDX générée ce qui permet un contrôle.

Supposons que l'utilisateur veuille répondre à la question : « Quelle est la surface du domaine d'étude 'Autre forêt' dont l'usage au sol est 'Accueil , Loisirs, Parc public, Habitat, Parc privé, Enclos' pour chaque zone phytogéographique ? »

Pour répondre à cette requête, en utilisant le Cube Panel, l'utilisateur sélectionne les membres des différentes dimensions qui doivent être affichées : « Autre Forêt » dans la dimension « Croisement Couverture x Utilisation x Taille Massif », « Accueil, Loisirs, Parc public, Habitat, Parc privé, Enclos » dans la dimension « Usage du sol » et le niveau « Zone phytogéographique » dans la dimension écologique.

L'interaction avec le serveur Mondrian se traduit par la création d'une requête MDX (figure 5-15) dont le résultat est visualisé simultanément sous forme cartographique (figure 5-16) et tabulaire (figure 5-17).

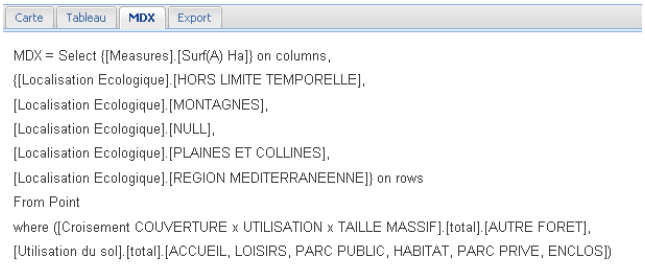


Figure 5-15. Requête MDX générée par GeoExt

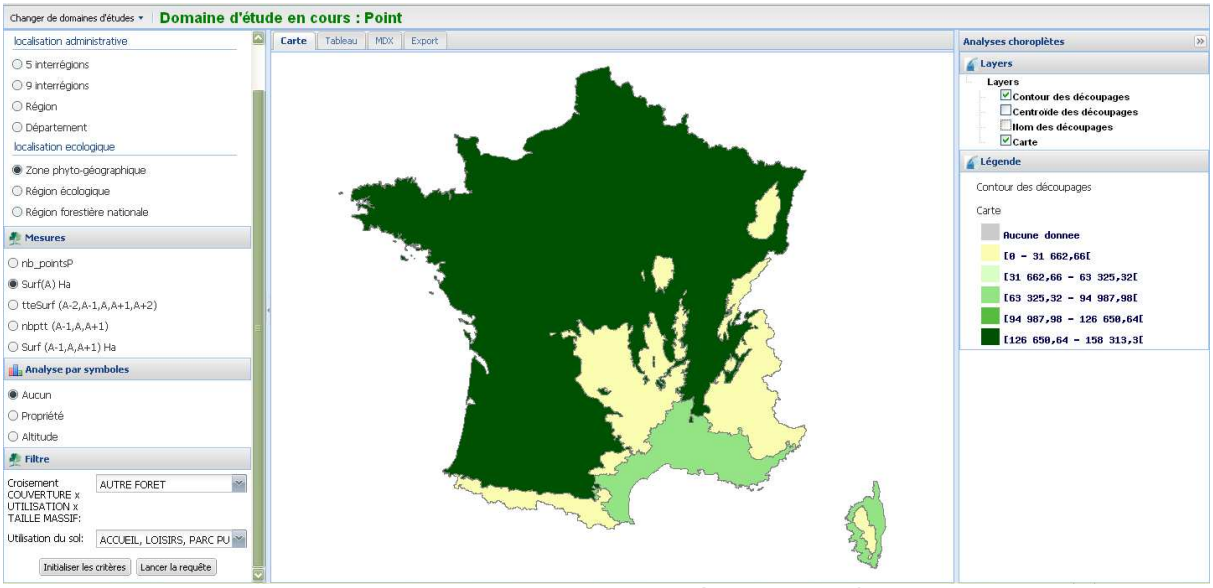


Figure 5-16. Visualisation cartographique du résultat de la requête MDX

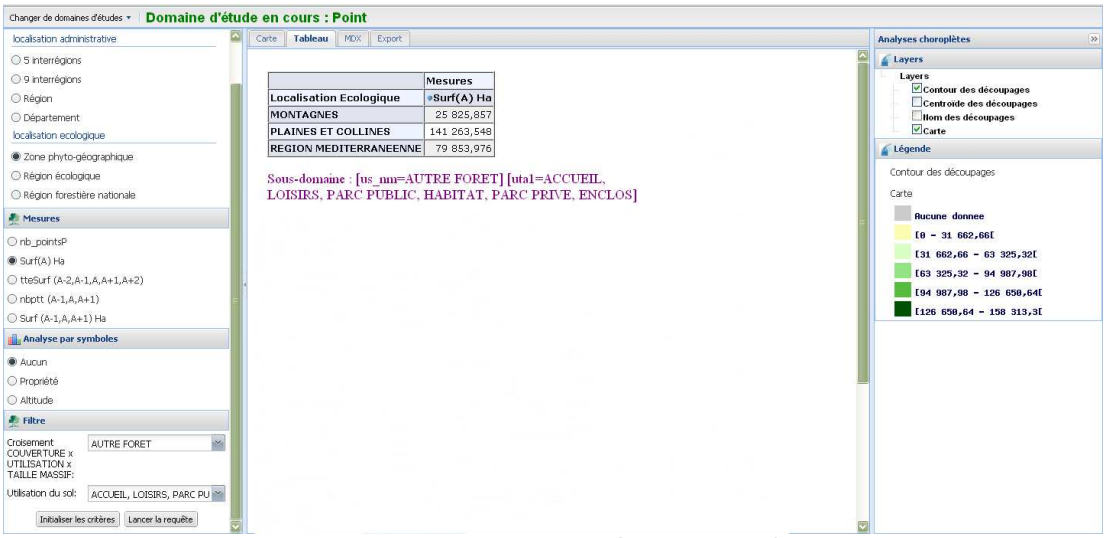


Figure 5-17. Visualisation tabulaire du résultat de la requête MDX

3 Synthèse

La première phase du travail réalisée et présentée aux directions (Cf. chapitre précédent), d'un commun accord la suite des développements ont été axés sur la dimension spatiale, avec pour objectif d'implémenter une solution web cartographique. Cette implémentation peut supporter facilement et rapidemment une exploration des données suivant l'approche multidimensionnelle utilisant des niveaux d'agrégations disponibles dans des affichages cartographiques aussi bien que dans des tableaux.

L'architecture dans son ensemble fonctionne et nous répondons aux critères décisionnels en pouvant effectuer des recherches exploratoires avec une restitution cartographique, les temps de réponse sont satisfaisants, l'interface conviviale bien qu'elle puisse être améliorée.

Voici une liste des améliorations et d'optimisations qui sont apparues au cours de la réalisation de ce prototype.

- Les performances semblent correctes. Toutefois des solutions existent afin d'améliorer les temps de réponse des requêtes et de l'affichage : des index spatiaux peuvent être implémentés en base et des caches sur le serveur cartographique.

- La mise en place de l'analyse supplémentaire par symbole est à terminer. Les résultats ventilés par cet axe supplémentaire sont en base, il suffit de gérer l'affichage des graphiques à l'aide du serveur MapServer.

- Nous nous sommes axés sur un certain type de sortie de résultat : une carte choroplèthe avec une analyse par symbole d'une variable supplémentaire, ce qui correspondrait dans une représentation tabulaire à un tableau à deux dimensions. Cela donc permet de visualiser par les cartes des résultats statistiques. Cependant en mettant en valeur d'autres dimensions nous aboutirions à d'autres fonctionnalités. Par exemple en tenant compte de la dimension temporelle et en ajoutant un affichage d'une carte par année de résultats, l'utilisateur pourrait visualiser des résultats en glissant d'une année à une autre.

- Actuellement le forage spatial passe par le formulaire, comme les coupes. En améliorant les fonctionnalités de navigation et en élaborant des outils dans une barre de navigation, l'exploration deviendrait plus dynamique. Par exemple en ajoutant un simple bouton de type (+), en cliquant dessus l'utilisateur déclencherait un forage spatial, i.e. un drill-down sur la dimension spatiale i.e.

une sélection d'un niveau inférieur. De même en ajoutant un bouton de type (-), l'évènement clic traduirait un drill-up.

- L'interface gagnerait en richesse en ajoutant des liens de navigation entre le tableau et la carte, des infos-bulles sur les régions géographiques de la carte indiquant le détail des données originales.

- La méthode de discrétisation est écrite en dur dans le code. Rendre paramétrable le choix des du nombre de classes, l'ajustement des différents seuils de la distribution et la méthode apporterait un plus dans l'autonomie de l'utilisateur et dans la facilité d'exécution.

- L'export en version PDF d'un rapport avec une déclinaison de carte à produire soulagerait l'utilisateur en l'épargnant d'exécuter sa liste de requête.

Conclusion

1 Rappel des objectifs

L'Inventaire Forestier National est en charge depuis cinquante ans de l'inventaire permanent des ressources forestières nationales. L'architecture mise en place afin de produire des résultats statistiques laisse voir quelques limites de performance, tant au niveau ergonomie qu'au niveau des performances pour un utilisateur souhaitant obtenir de nombreux résultats en ligne.

Les entrepôts de données associés à des outils d'analyse On Line Analytical Processing, OLAP représentent une solution effective pour l'informatique décisionnelle. Les données qualitatives des inventaires sont organisées dans des hypercubes en axes d'analyses appelés dimensions. Les sujets d'analyse, appelés tables de fait, sont caractérisées par les données quantitatives appelées mesures. Les résultats issus de cette analyse sont calculés à l'aide de fonctions d'agrégations selon les différentes granularités définies par le schéma hiérarchique des dimensions.

Aujourd'hui les outils décisionnels sont beaucoup plus puissants que les systèmes relationnels dotés d'extensions multidimensionnelles (opérateurs CUBE, ROLLUP, …). Par exemple le langage MDX permet de créer ces propres fonctions d'agrégations appliquées aux mesures. Le calcul des résultats d'inventaire proviennent de calculs complexes, issus de la statistique du sondage. Les estimations des variables sont additives suivant les axes des principales dimensions.

Le premier objectif visait à mettre en place un système d'entrepôt de données afin de servir de base à la production de résultats. Par la même cette action a permis de combler une lacune dans le processus d'analyse : naviguer dans le cube multidimensionnel, visualiser les mesures pour des ensembles de membres à des niveaux de granularité sélectionnés par l'utilisateur et acquérir des interfaces orientées navigation (tableau de bord).

2 Synthèse

Le premier trimestre a été consacré principalement à appréhender les techniques liées aux entrepôts de données et à les mettre en œuvre. Le développement de cet objectif a été réalisé et terminé au deux-tiers de mon stage. Le choix de l'architecture, ainsi que des logiciels open source déployés ont été décidés très rapidement. Dans la phase de recherche qui a précédé le début du stage en janvier 2009, j'avais participé à un séminaire, organisé par la société Smile[34], dont le sujet « Décisionnel – Solutions Open Source » faisait un état des lieux des meilleures solutions avec démonstration interactive. La solution Mondrian/JPivot correspondait exactement aux besoins.

Dès que les données des modélisations physiques ont été chargées, les modèles logiques implémentés sur le serveur Mondrian et l'application pivot en place, des présentations ont été réalisées dans les différentes directions (Direction générale, direction technique et direction de la valorisation) courant avril 2009 afin de présenter l'état d'avancement, valider le travail effectué et évaluer quelles étaient les directions à explorer pour répondre aux mieux aux possibles utilisations de ces techniques au sein de l'IFN.

Le bilan de la première phase s'avère très positif. L'approche exploratoire répond à de multiples attentes internes et externes. Le premier objectif est atteint.

[34] Smile www.smile.fr

A ce stade de la réalisation de multiples pistes se présentaient. Une éventualité aurait été de développer les calculs de coefficient de variation et des variances. La modélisation de l'entrepôt de donnée réussie pour le calcul des estimateurs, il est probable que les calculs des intervalles d'erreur, cependant beaucoup plus complexes, pourront être eux aussi intégrés aux analyses OLAP de l'entrepôt de données. Les mesures de l'IFN ainsi que leur coefficient de variation sont des mesures exhaustives. Les mettre en place dans l'entrepôt de données demanderait un investissement en temps trop important et limitera les autres champs d'investigation demandés. Les calculs des variances des estimateurs n'ont donc pas été mis en place dans cette étude pour l'instant.

L'objectif suivant qui convenait aux directions, consistait à valoriser les dimensions spatiales dans le processus de décision mis en place précédemment. C'est dans cette direction que se positionnent donc les travaux de la deuxième phase. Ils concernent l'introduction de l'information spatiale dans OLAP. Sans outils open source aujourd'hui sur le marché, l'enjeu a été d'implémenter une application qui puisse permettre une extraction interactive de connaissances géographiques, les visualiser via une production de cartes sur le web et d'explorer les cubes de données par des forages spatiaux.

Ce développement a nécessité trois bons mois de travail de mai 2009 à juillet 2009. Tout d'abord il a fallu apprendre à maîtriser la librairie GeoExt, apprivoiser le serveur MapServer ainsi que le client OpenLayers et utiliser les classes d'Olap4j. A l'issue de cette période d'apprentissage, la réalisation de ce prototype a été assez rapide et terminée au milieu de l'été 2009. Il reste encore à la présenter aux directions.

Le bilan de cette réalisation s'avère positif quant au choix de l'architecture choisie : dialogue entre les serveurs ROLAP et cartographique, visualisation synchrone des cartes et des tableaux, formulaire ajax interactif, zoom sur les différents niveaux des dimensions géographiques, filtres sur certaines dimensions, performance et temps de réponse correct.

Dans son ensemble, la mise en place de l'entrepôt de données est un terreau fertile pour les idées qui ne manquent pas de pousser. Le langage MDX permet d'effectuer des calculs complexes. L'emploi de ce langage dans les calculs des estimateurs, m'a permis aujourd'hui d'atteindre une certaine maturité sur les calculs d'agrégats qui pourra combler l'abscence de calcul des intervalles de confiance. D'autres domaines comme les tableaux de bord spatiaux, la data mining spatial, laissent entrevoir encore beaucoup de perspectives de mises en application du géodécisionnel.

Ressources

1 Bibliographie

[Beguin] Michèle Béguin, Denise Pumain. La représentation des données géographiques.

[Bimonte] Sandro Bimonte. Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne : de la modélisation à la visualisation. 2007.

[Davis] Scott Davis. Gis for Web Developpers. Adding Where to Your Web Applications. 2007

[Frederick] Shea Frederick, Colin Ramsay, Steeve Cutter Blades. Learning Est JS. 2008.

[Guerrero] Edgard Benitez Guerrero, Christine Collet, Michel E. Adiba. Entrepôts de données : Synthèse et Analyse. 1999.

[IFN 2007] La forêt française – Les résultats issus des campagnes d'inventaire 2005 et 2006.

[Kimball] Ralph Kimball et Margy Ross. Entrepôts de données – Guide pratique de modélisation dimensionnelle.

[Miquel] Maryvonne Miquel, Yvan Bédard, Alexandre Brisebois. Conception d'entrepôts de données géospatiales à partir de sources hétérogènes. Exemple d'application en foresterie.

[Morin] Annie Morin, Patrick Bosc, Georges Hebrail, Ludovic Lebart. Bases de données et statistique.

[Plumejeaud] Christine Plumejeaud. Acquisition de données et cartes de potential pour l'analyse spatiale. 2007.

[Shekhar] S. Shekhar, C. T. Lu, X. Tan, S. Chawla, R. R. Vatsavai. Map Cube: A Visualization Tool for Spatial Data Warehouses. 2001.

[SMILE] : Nicolas Richeton, Bad Chentouf. Livre blanc – Décisionnel solutions open source. Edition 2010.

[Spofford] : George Spofford, Sivakumar Harinath, Christopher Webb, Dylan Hai Huang, Francesco Civardi. MDX Solutions With Microsoft SQL Server Analysiis Services 2005 and Hyperion Essbase. 2006.

2 Webographie

[@Badard] http://geosoa.scg.ulaval.ca/fr/index.php

[@Bedard] http://sirs.scg.ulaval.ca/yvanbedard/

[@JPivot] http://jpivot.sourceforge.net/

[@Kaouich] http://eric.univ-lyon2.fr/~kaouiche/inf9002/index.html

[@MapServer] http://mapserver.org/

[@Mondrian] http://mondrian.pentaho.com/

[@OpenLayers] http://openlayers.org/

[@PostgreSQL] http://www.postgresql.org/

[@Postgis] http://postgis.refractions.net/

Annexes

1 Schéma du cube Point

```
<?xml version="1.0" encoding="UTF-8"?>

    <Schema name="DW5">

        <Dimension name="Année">
            <Hierarchy hasAll="false" primaryKey="n_str" defaultMember="[Année].[2007]">
            <Table name="strate"/>
            <Level name="Année" column="annee" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>


        <Dimension name="Stratification">
            <Hierarchy hasAll="false" primaryKey="n_str"
defaultMember="[Stratification].[Stratification Ph2 par DEP, types FAO, PROpriete et Ph1]">
            <Table name="strate"/>
            <Level name="stratification" column="stratification" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>


        <Dimension name="Localisation Administrative">
            <Hierarchy hasAll="true" allMemberName="France entière" primaryKey="_dep">
                <Table name="dep"/>
                <Level name="5 interrégions" column="dep_rad2_libelle" uniqueMembers="true">
                    <Property name="geometrie" column="dep_rad2_geom" />
                </Level>
                <Level name="9 interrégions" column="dep_rad3_libelle" uniqueMembers="true">
                    <Property name="geometrie" column="dep_rad3_geom" />
                </Level>
                <Level name="Région" column="dep_ra_libelle" uniqueMembers="true">
                    <Property name="geometrie" column="dep_ra_geom" />
                </Level>
                <Level name="Département" column="dep_dp_libelle" uniqueMembers="true">
                    <Property name="geometrie" column="dep_dp_geom" />
                </Level>
            </Hierarchy>
        </Dimension>

        <Dimension name="Localisation Ecologique">
               <Hierarchy hasAll="true" allMemberName="France entière" primaryKey="_regn">
                   <Table name="regn"/>
                   <Level name="Zone phyto-géographique" column="regn_zp_libelle"
uniqueMembers="true">
                       <Property name="geometrie" column="regn_zp_geom" />
                   </Level>
                   <Level name="Région écologique" column="regn_zpg0_libelle"
uniqueMembers="true">
                       <Property name="geometrie" column="regn_zpg0_geom" />
                   </Level>
                   <Level name="Région forestière nationale" column="regn_rf_libelle"
uniqueMembers="true">
                       <Property name="geometrie" column="regn_rf_geom" />
                   </Level>
               </Hierarchy>
        </Dimension>


        <Dimension name="Propriété">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_pro_nm">
                <Table name="pro_nm"/>
                <Level name="Statut forestier" column="pro_nm_sns_libelle"
uniqueMembers="true">
                    <Property name="symbole" column="sns_libelle" />
                </Level>
                <Level name="Propriété" column="pro_nm_pf_libelle" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>
```

```
        <Dimension name="Altitude">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_clz">
            <Table name="clz"/>
            <Level name="Classe d'altitude" column="clz_clz_libelle" uniqueMembers="true">
                <Property name="symbole" column="sns_libelle" />
            </Level>
            </Hierarchy>
        </Dimension>

        <Dimension name="Croisement COUVERTURE x UTILISATION x TAILLE MASSIF">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_us_nm">
            <Table name="us_nm" />
            <Level name="us_nm" column="us_nm_csuttm_libelle" uniqueMembers="true"
ordinalColumn="us_nm_csuttm_ordre">
                <Property name="filtre" column="resfeuil_r_libelle" />
            </Level>
            </Hierarchy>
        </Dimension>

        <Dimension name="Information sur la catégorie du point">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_info">
            <Table name="info"/>
            <Level name="Info" column="info_info_libelle" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>

        <Dimension name="Leve de terrain">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_leve">
            <Table name="leve"/>
            <Level name="Leve" column="leve_leve_libelle" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>

        <Dimension name="Taille de formation">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_tm2_i01">
            <Table name="tm2_i01"/>
            <Level name="tm2_i01" column="tm2_i01_tm_libelle" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>

        <Dimension name="Utilisation du sol">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_uta">
            <Table name="uta"/>
            <Level name="uta1" column="uta1_uta_libelle" uniqueMembers="true">
                <Property name="filtre" column="resfeuil_r_libelle" />
            </Level>
            </Hierarchy>
        </Dimension>

        <Dimension name="Deuxième utilisation du sol">
            <Hierarchy hasAll="true" allMemberName="total" primaryKey="_uta">
            <Table name="uta"/>
            <Level name="uta2" column="uta2_uta_libelle" uniqueMembers="true"/>
            </Hierarchy>
        </Dimension>


    <!--  ============================  -->
    <!--            POINT               -->
    <!--  ============================  -->

    <Cube name="Point" defaultMeasure="nb_pointsP">

        <Table name="cube_point" />

        <DimensionUsage name="Année" source="Année" foreignKey="n_str" />
        <DimensionUsage name="Stratification" source="Stratification" foreignKey="n_str" />

        <DimensionUsage name="Localisation Administrative" source="Localisation
Administrative" foreignKey="dep" />
        <DimensionUsage name="Localisation Ecologique" source="Localisation Ecologique"
foreignKey="regn" />
        <DimensionUsage name="Propriété" source="Propriété" foreignKey="pro_nm" />

        <DimensionUsage name="Altitude" source="Altitude" foreignKey="clz" />
        <DimensionUsage name="Leve de terrain" source="Leve de terrain" foreignKey="leve" />
```

```
        <DimensionUsage name="Information sur la catégorie du point" source="Information sur
la catégorie du point" foreignKey="info" />

        <DimensionUsage name="Croisement COUVERTURE x UTILISATION x TAILLE MASSIF"
source="Croisement COUVERTURE x UTILISATION x TAILLE MASSIF" foreignKey="us_nm" />
        <DimensionUsage name="Taille de formation" source="Taille de formation"
foreignKey="tm2_i01" />
        <DimensionUsage name="Utilisation du sol" source="Utilisation du sol" foreignKey="uta"
/>
        <DimensionUsage name="Deuxième utilisation du sol" source="Deuxième utilisation du
sol" foreignKey="uta" />


        <!-- Mesure "Nombre de placette" -->
        <Measure name="nb_pointsP" column="point_nb" aggregator="sum" visible="true" />

        <Measure name="Surf(A) Ha" aggregator="sum" visible="true">
            <MeasureExpression>
                      <SQL dialect="generic">
                 cube_point.point_poids * cube_point.point_sk / cube_point.point_pk / 1E4
                      </SQL>
            </MeasureExpression>
        </Measure>

        <CalculatedMember name="tteSurf (A-1,A,A+1)" dimension="Measures" visible="false">
            <Formula>( ([Année].currentmember,[Measures].[Surf(A) Ha]) +
                  ([Année].currentmember.prevmember,[Measures].[Surf(A) Ha]) +
                  ([Année].currentmember.nextmember,[Measures].[Surf(A) Ha])  ) / 3</Formula>
            <CalculatedMemberProperty name="SOLVE_ORDER" value="1" />
        </CalculatedMember>

        <CalculatedMember name="tteSurf (A-2,A-1,A,A+1,A+2)" dimension="Measures"
visible="true">
            <Formula>( ([Année].currentmember,[Measures].[Surf(A) Ha]) +
                  ([Année].currentmember.prevmember,[Measures].[Surf(A) Ha]) +
                  ([Année].currentmember.prevmember.prevmember,[Measures].[Surf(A) Ha]) +
                  ([Année].currentmember.nextmember.nextmember,[Measures].[Surf(A) Ha]) +
                  ([Année].currentmember.nextmember,[Measures].[Surf(A) Ha])  ) / 5</Formula>
            <CalculatedMemberProperty name="SOLVE_ORDER" value="1" />
        </CalculatedMember>

        <CalculatedMember name="neutreSurf (A-1,A,A+1)" dimension="Measures" visible="false">
            <Formula> iif([Année].currentmember is [Année].[2005],
                          null,
                          iif([Année].currentmember is [Année].[2009],
                              null,
                              [Measures].[tteSurf (A-1,A,A+1)]))</Formula>
            <CalculatedMemberProperty name="SOLVE_ORDER" value="2" />
        </CalculatedMember>

        <CalculatedMember name="nbptt (A-1,A,A+1)" dimension="Measures" visible="true">
            <Formula>( ([Année].currentmember,[Measures].[nb_pointsP]) +
                       ([Année].currentmember.prevmember,[Measures].[nb_pointsP]) +
                       ([Année].currentmember.nextmember,[Measures].[nb_pointsP])  )</Formula>
            <CalculatedMemberProperty name="SOLVE_ORDER" value="3" />
        </CalculatedMember>

        <CalculatedMember name="Surf (A-1,A,A+1) Ha" dimension="Measures">
            <Formula>[Measures].[neutreSurf (A-1,A,A+1)]</Formula>
            <CalculatedMemberProperty name="FORMAT_STRING"
                     expression="IIf(([Measures].[nbptt (A-1,A,A+1)] > 50.0),
                                     '|#,##0|style=green',
                                     IIf(([Measures].[nbptt (A-1,A,A+1)] > 10.0),
                                          '|#,##0|style=yellow',
                                          '|#,##0|style=red'))" />
            <CalculatedMemberProperty name="SOLVE_ORDER" value="4" />
        </CalculatedMember>


    </Cube>

</Schema>
```

2 Exemple d'utilisation de la librairie Olap4j

Exemple d'appel aux objets de l'API Olap4j afin d'éxecuter une requête MDX, de la parser et d'ajouter les résultats en base.

```
// Execute a statement.
OlapStatement statement = olapConnection.createStatement();
CellSet result = statement.executeOlapQuery(mdxRequest);

List<CellSetAxis> cellSetAxes = result.getAxes();
ArrayList<Double> listeValeur = new ArrayList <Double> ();
CellSetAxis columnsAxis = cellSetAxes.get(Axis.COLUMNS.ordinal()-1);
CellSetAxis rowsAxis = cellSetAxes.get(Axis.ROWS.axisOrdinal()-1);
int cellOrdinal = 0;
for (Position rowPosition : rowsAxis.getPositions()) {

    for (Position columnPosition : columnsAxis.getPositions()) {

        // Access the cell via its ordinal. The ordinal is kept in step
        // because we increment the ordinal once for each row and
        // column.
        Cell cell = result.getCell(cellOrdinal);

        List<Integer> coordList = result.ordinalToCoordinates(cellOrdinal);
        assert coordList.get(0) == rowPosition.getOrdinal();
        assert coordList.get(1) == columnPosition.getOrdinal();

        ++cellOrdinal;

        // pour chacune des mesures
        for (Member member : columnPosition.getMembers()) {

            for(Property prop : member.getProperties()){

                if(prop.getName().trim().equals("geometrie")){

                    PGgeometry geom  = (PGgeometry)member.getPropertyValue(prop);
                    if(geom != null){
                        String insertLayer = " Insert into spatialuser.usergeom ( "
                          + "         sessionid, nom_geom, resultat, codecouleur, the_geom )  "
                          + "  values ('" + sessionID + "', "
                          + "         '" + member.getName().trim().replace("'", "''") + "', "
                          + "         " + cell.getValue() + ", "
                          + "         '000', "
                          + "         '" + geom.toString() + "') ";
                        Statement stm = maConnection.createStatement();
                        stm.executeUpdate(insertLayer);
                        stm.close();
                        listeValeur.add((Double)cell.getValue());
                    }
                }
            }
        }
    }
}
```