Université de Versailles

-

Thèse de Doctorat
de Méthodes Informatiques
soutenue le 12 décembre 1997

-

Institut Géographique National (IGN)
Laboratoire COGIT

# Processus d'intégration et d'appariement de Bases de Données Géographiques

## Application à une base de données routières multi-échelles

**Thomas Devogele**

**Jury :**

| | |
|---|---|
| André Frank | : Rapporteur |
| Geneviève Jomier | : Rapporteur |
| Eric Simon | : Directeur de thèse |
| Mokrane Bouzeghoub | : Président |
| Robert Laurini | : Examinateur |
| Laurent Raynal | : Examinateur |
| François Salgé | : Examinateur |
| Stefano Spaccapietra | : Examinateur invité |

*A GIS brings information together, it unifies and integrates that information. It makes available information to which no one had access before, and places old information in a new context. It often brings together information which either was not or could not be brought together previously.*

*J. Dangermond 1989 (ESRI)*

*The organizational impact of GIS technology*

*ARC News Summer : 25-6*

# Remerciements

Une thèse n'est jamais un travail individuel, elle s'inscrit dans une démarche globale de recherche qui bénéficie des travaux déjà réalisés, est en relation avec les travaux en cours et sera reprise par la suite.

Cette thèse doit donc énormément à Laurent Raynal qui est l'initiateur des travaux en multi-représentations au COGIT. Je le remercie amicalement pour son encadrement constant tout en douceur et ses nombreux conseils (y compris pendant ses nouvelles fonctions).

Ma gratitude va aussi à la hiérarchie de l'IGN (Jacques Poulain, Hervé Le Men, Serge Motet, Sylvie Lamy et François Chirié) pour avoir permis cette thèse et lui avoir donné les moyens de se développer dans des conditions favorables.

Je tiens à remercier particulièrement Jean Philippe Lagrange pour m'avoir fait partager ses connaissances encyclopédiques aussi bien en généralisation qu'en BDG, pour sa part active au début de la thèse et sa relecture qui a permis de compléter et d'affiner cette thèse.

Je dois énormément à Stefano Spaccapietra et Christine Parent qui m'ont partagé leurs connaissances en intégration des BD. Ils ont largement contribués à la qualité scientifique de cette thèse par leur visite et leur collaboration à ses recherches. Je désire leur exprimer toute ma reconnaissance.

Je remercie également Sabine Timpf pour sa visite et notre travail en commun sur les applications multi-représentations qui m'a fourni une vision complémentaire de ce problème.

Je remercie mon directeur de thèse Eric Simon qui a su diriger mes recherches en insistant sur les grandes lignes de cette thèse et en évitant les dispersions.

Je tiens à remercier tous les membres de mon jury qui se sont intéressés à ce travail. Plus particulièrement, je tiens à exprimer ma grande reconnaissance aux rapporteurs ; à Geneviève Jomier, qui s'est toujours préoccupée de mes recherches et m'a souvent encouragé, et à André Frank dont les recherches en multi-représentations ont été un support pour mes travaux et qui malgré la distance Paris - Vienne a accepté la tâche de rapporteur. Je remercie aussi, Georges Gardarin ainsi que Mokrane Bouzeghoub qui m'ont accueilli au sein du laboratoire PRISM de l'université de Versailles, Robert Laurini dont les travaux en interopérabilité ont été une des sources principales de ma taxonomie des conflits, et enfin, François Salgé qui est un des pères de l'intégration de BDG entre les différents instituts cartographiques Européens.

Cette thèse s'est déroulée au laboratoire COGIT de l'IGN je tiens donc à remercier tous les membres du COGIT et du secrétariat de la DT pour leur compétence, l'ambiance de travail dans la joie et l'entraide qu'ils savent si bien entretenir. Un grand merci à Cécile Lemarié, Thierry Badard et François Vauglin pour leurs critiques constructives, leur relecture et leur collaboration active à ce travail. Je remercie aussi les stagiaires qui se sont succédés (Jenny Trevisan, Roger Areia, Paolo Fereira, Patrick Branly et Odile Ousset) qui m'ont assisté pour développer le prototype. Merci encore à Pascale Monier et Jean-François Hangouët pour leurs relectures.

Je tiens aussi à encourager Matthieu Castagnet et Thierry Badard qui reprendront ces travaux dans le cadre de leur thèse pour les amener vers d'autres horizons.

Je désire terminer, en remerciant ma femme et mes deux filles ; ces trois sources de joie indispensable à l'équilibre de ce travail de thèse.

# Résumé

Les phénomènes du monde réel sont actuellement représentés dans les Bases de Données Géographiques (BDG) sous différentes formes (suivant les modèles géographiques, suivant les points de vue utilisateur et/ou suivant les systèmes). La réutilisation de telles BDG nécessite dès lors un processus d'intégration pour éliminer les parties redondantes et regrouper les parties complémentaires. Ce processus d'intégration est nécessaire étant donné le coût d'acquisition des données géographiques (de cette manière des saisies d'information sont évitées) et permet de fédérer l'information provenant de différentes sources. L'intégration est donc au cœur du problème d'interopérabilité entre BDG.

A l'issue de celle-ci, plusieurs représentations de phénomènes du monde réel sont disponibles selon des points de vue différents et des échelles distinctes. Ces représentations multiples sont nécessaires pour des applications très diverses : cartographie électronique multi-échelle, propagation des mises à jour, aide à la navigation.

L'objectif de cette thèse consiste donc à définir un processus d'intégration de BDG sur un seul site, le processus étant limité aux données en mode vecteur à deux dimensions. Il propose l'extension d'un processus d'intégration classique à trois phases [Spaccapietra et al. 92] (pré-intégration, déclaration des correspondances, intégration). L'extension est fondée sur une taxonomie des conflits d'intégration entre BDG et sur l'ajout d'un processus d'appariement géométrique et topologique. Ce processus a été mis en œuvre sur les trois principales bases de données de l'IGN (BD TOPO®, BD CARTO® et GEOROUTE®) pour le thème routier dans la région de Lagny (environ 900 km de tronçons routiers).

Etant donnée la complexité des phénomènes géographiques, plusieurs interprétations et donc plusieurs modélisations des phénomènes peuvent être définies. La taxonomie des conflits d'intégration de BDG effectue une structuration de ces différences : conflits de définition de classe (conflits de classification, conflits de fragmentation, conflits de spécification), conflits d'hétérogénéité, conflit de description,…Six catégories de conflits ont été traitées dans le processus d'intégration.

Certains conflits sont pris en compte dans la phase de pré-intégration. D'autres font l'objet d'un traitement spécifique : extension du langage de déclaration des correspondances, ajout d'opérations de résolution de ce conflit. De plus, la phase d'intégration doit suivre une stratégie. Cette stratégie détermine le choix des opérations et fixe l'objectif de l'intégration. Au vu de nos bases d'expérimentations, deux stratégies d'intégration (et leurs opérations d'intégration associées) sont présentées.

Le processus d'appariement consiste à identifier les données représentant le même phénomène du monde réel et permet le regroupement d'informations. Cette étape est précieuse car elle enrichit les BDG d'opérations inter-représentations, opérations nécessaires aux applications multi-représentations.

Un processus d'appariement a été développé pour les données de types routières à différentes échelles. Les résultats obtenus font apparaître un taux de correspondance de l'ordre de 90 %. Un processus générique en a été déduit afin de guider la conception des processus d'appariement concernant d'autres types de données.

Cette thèse apporte donc un cadre général et détaillé pour les intégrations de BDG et contribue ainsi à l'essor d'applications multi-représentations et de l'interopérabilité entre les BDG en adaptant ces processus à des BDG réparties sur un réseau.

**Mots clés** : base de données géographiques, système d'information géographique, intégration, appariement, multi-représentations, multi-échelles, conflit d'intégration, interopérabilité, réutilisation.

# Abstract

Phenomena of the real world are described in a variety of forms in current geographic data bases (GDBs) : geographic data models, users' points of view, systems are different. Concurrently re-using GDBs thus requires an **integration process** both to eliminate duplicates and to regroup complements. Integration makes it possible to federate data from different sources while cutting down acquisition costs (new data captures are avoided) ; it is a crucial issue for interoperability between GDBs.

After integration, several representations of real world phenomena are available, with distinct points of view and at different scales. These **multiple representations** are useful and even necessary for a wide range of applications, such as multi-scale cartography, update propagation, aided navigation etc.

The aim of the thesis is to devise an integration process on 2-D, vector data of a single-site GDB. It is modelled as an extended classical three-stepped integration process (schema preparation, investigation for correspondences, integration) [Spaccapietra et al. 92]. The extension involves a taxonomy of integration conflicts between GDBs and a process for joint geometric / topologic data matching. The integration process has been applied from IGN's three main data bases (BD TOPO®, BD CARTO® and GEOROUTE®) on the Lagny area (900 km of roads network).

Given the complexity of the real world's phenomena, several versions and as many models may be defined. The differences are structured in the **taxonomy of integration conflicts** : conflicts come as definition conflicts (classification conflicts, fragmentation conflicts, specification conflicts), heterogeneity conflicts, description conflicts... Altogether, six conflict categories have been used for the integration process.

Some conflicts are taken into account at the stage of schema preparation. Other conflicts require specific treatment : extension of the language of correspondence declarations and operations to solve the conflicts. Indeed, integration requires a strategy. The **strategy** makes the choice of operations to perform and fixes the goal of the integration process. For the databases of the experiments, two strategies and their associated operations are shown.

The **data matching process** consists in identifying sets of data representing the same real world phenomenon and allows to regroup data. This step is precious for it enhances GDBs with inter-representation operations that are useful for multi-representation applications.

A data matching process has been developed for road data at different scales, with 90% of the results correct. Henceforth, a generic process has been inferred to help designing matching processes on other kinds of data.

The thesis, describing a generic and detailed framework for the integration of GDBs, contributes to the development not only of multi-representation applications but also of interoperability between GDBs , once the processes are adapted to network distributed GDBs.



**Key words** : geographic database, geographic information system, integration, data matching, multi-representation, multi-scale, integration conflict, interoperability, reusing

# Sommaire

# Figures

# Tableaux

# 1. Introduction

## 1.1 Contexte : SIG et information géographique

Certaines activités humaines doivent être localisées dans l'espace, évaluer des distances, mesurer des surfaces. D'autres ont besoin d'information sur le paysage, les objets et leur forme [Souquiere 94]. La **carte** a servi (et continue de servir) de moyen de communication privilégié pour transmettre de tels messages ; c'est une vue, une représentation particulière d'un espace géographique adaptée à une utilisation précise, avec des informations différentes.

Pour composer une carte, le rôle de **l'échelle** (rapport entre une mesure de distance sur la carte et une même mesure sur le terrain) est majeur. Elle détermine la taille de la carte, la sélection des objets et leur représentation par leur emprise ou un symbole en fonction de leur taille.

Actuellement, l'introduction des **systèmes d'information géographique** (SIG) ouvre de nouvelles possibilités. Ils doivent répondre à des applications qui dépassent de loin la cartographie automatique, telles l'aide à la décision, à la gestion du territoire, à la planification dans des domaines très variés (urbanisme, environnement, transport, …). Ils permettent le stockage, la manipulation, l'interrogation, le partage, la diffusion, la restitution à l'écran et sur papier d'informations à composante géographique. Dans cet objectif, un SIG regroupe :

- un système de gestion de bases de données (SGBD) dédié à la gestion de données géographiques,
- des outils d'analyse spatiale appropriés au domaine.

Les **données géographiques** sont des abstractions des **phénomènes du monde réel**[1]. Elles peuvent être **localisées** (géo-référencées) dans un espace à deux ou trois dimensions. Une donnée géographique est décrite par sa **géométrie** (ensemble de localisation dans l'espace), ses données **sémantiques ou descriptives** (nom, type de l'entité, …), des **relations topologiques** avec les autres entités géographiques (inclusion, adjacence, …) et des **relations classiques** (par exemple, une entité département a pour préfecture une entité ville). L'abstraction des phénomènes du monde réel est d'autant plus simplifiée que « **l'échelle** » est petite. Formellement, une **base de données géographiques** (**BDG**) n'inclut pas la notion d'échelle ; [Goodchild 91] [Müller et al. 95] parlent plutôt, et à juste titre :

- **de précision** : le degré de détail dans les mesures, par exemple, les mesures peuvent être en mètre ou en centimètre,
- **d'exactitude** : l'écart entre la mesure stockée dans la BDG et une mesure parfaite, qui ne serait entachée d'aucune erreur,
- **de résolution géométrique** : la taille du plus petit objet représentable,
- **de résolution sémantique** : le niveau de détail dans la description sémantique des entités.

Il semble donc judicieux de rapprocher la détermination de ces quatre notions pour une base de données géographiques à la notion d'échelle communément admise pour une carte.

[1] Un phénomène du monde réel est une réalité qui se manifeste à la conscience, que ce soit par l'intermédiaire des sens ou non. Il constitue la réalité première [Gouvernement du Québec 96]. Par exemple, la route nationale 7, les bâtiments du 2 avenue Pasteur, la limite de la ville de Paris.

Les SGBD géographiques ont donc des caractéristiques qui leur sont propres. Ils doivent :

- **modéliser la localisation** de leurs données. Cette modélisation n'est pas possible avec les types classiques des SGBD (entier, réel, caractère,…). Des **types complexes** doivent donc être définis.
- **visualiser les données géographiques à l'écran**. Pour les données classiques, il suffit d'afficher les valeurs des attributs. Par contre, pour les données géographiques, les valeurs de leur géométrie sont obscures pour l'utilisateur ; elles doivent être converties en utilisant des symboles graphiques de manière à les rendre compréhensibles.
- **prendre en compte la qualité des données**. Les données géographiques sont issues de **sources** de qualités différentes (relevés topographiques, photographies aériennes, …) ; et les limites des données géographiques (forêt, chemin, …) sont souvent vagues. Les utilisateurs ont donc besoin de connaître la qualité des données géographiques contenues dans leur base [David 97] [Vauglin 97], afin d'en tenir compte lors de leurs requêtes.

## 1.2 Les SGBD géographiques multi-représentations

Les SGBD géographiques peuvent aussi représenter les données d'une même zone suivant différents points de vue : ils sont alors appelés **SGBD géographiques multi-représentations** [Brugger et al. 89]. Une sous-classe de ces SGBD sont les **SGBD géographiques multi-échelles**, qui permettent de représenter les données d'une même zone suivant différentes échelles.

Les représentations multiples sont un des problèmes clé dans le domaine des SIG [Brugger et al. 89]. En effet, pour l'analyse spatiale et la recherche d'itinéraire, une méthode de raisonnement avec différents niveaux de détails, de résolution, de précision, est une bonne démarche [Mark 89].

Deux méthodes sont envisageables pour concevoir une BD multi-représentation :

- la première méthode consiste à avoir une seule base de données de référence : la base de données la plus détaillée, et un processus pour générer les autres représentations (1.3),
- la deuxième méthode que nous avons retenue, consiste à définir un processus d'intégration des représentations géographiques existantes (1.4).

## 1.3 Limites actuelles de la généralisation de représentation géographique

Le **changement de représentation,** et plus particulièrement la généralisation de représentations géographiques, sont des problèmes complexes. Actuellement, les seuls changements réalisables sont des modifications « simples », telles les suppressions d'entités ou la reclassification. Les modifications qui touchent à la géométrie ou qui créent de nouvelles entités à partir des entités existantes ne sont quasiment pas traitées par les SIG du marché. La difficulté majeure réside en la généralisation de représentations géographiques.

### *1.3.1 La généralisation de représentation géographique*

La **généralisation** [Brassel et Weibel 88] [Lagrange et Ruas 94] [Ruas et Lagrange 95] de représentation géographique consiste à modifier les données, afin d'obtenir une représentation plus simple et plus abstraite. Le but de la généralisation est donc de réduire le nombre de données transmises en conservant autant que possible l'essentiel de l'information qu'elles véhiculent. Elle englobe une simplification descriptive (ou sémantique) et géométrique.

La généralisation de représentation géographique semble difficilement réalisable à l'heure actuelle car :

- il n'existe pas tous les outils permettant de généraliser automatiquement les représentations les moins détaillées à partir de la représentation la plus détaillée. En effet, les modifications en généralisation sont complexes, variées, très imbriquées et pas toujours automatisables [Müller et al. 95].
- une généralisation entièrement interactive est trop longue et pénible. Elle semble donc inadéquate pour un utilisateur final [Müller 91].
- toutes les informations nécessaires à la représentation généralisée ne sont pas présentes dans la représentation de référence car souvent le contexte est modifié.

### *1.3.2 Les structures de stockage*

Pour fournir à l'utilisateur des représentations selon différents critères (échelles, points de vue, …), une solution à court terme est de **maintenir des représentations multiples** des mêmes phénomènes du monde réel à des échelles différentes dans une même BDG [Kidner 96].

Les représentations obtenues à partir d'un processus de généralisation interactif, sont matérialisées dans des **structures de stockage arborescentes** ([Jones 91], [Kidner et Jones 94], [Ware et C. Jones 92], [Ware 94], [van Oosterom et Schenkelaars 91], [van Oosterom 95], [Timpf et Frank 95] [Rigaux 94]). Or les relations du monde réel ne sont pas toujours hiérarchiques quand l'échelle varie.

Idéalement, pour autoriser la simplification descriptive et les changements de représentations, il faudrait coupler ces structures avec les mécanismes de vues [Günther 89] [Abel et al. 94 a] [Abel et al. 94 b]. Le **mécanisme de vues** [Souza dos Santos 94] est employé pour donner une interface adéquate à chaque utilisateur de la base de données.

Cependant, ces deux approches n'ont pas été reliées : les transformations (géométrique et sémantique) ne peuvent pas être prises en compte par les mécanismes de vue classiques, inversement, les structures de stockage ne permettent pas la gestion de plusieurs points de vues.

En conclusion, toutes ces structures sont **limitées** car elles traitent soit des changements sémantiques, soit des changements géométriques. Or, les changements dus à la diminution de l'échelle combinent les changements sémantiques et géométriques. Le seul modèle autorisant quelques changements sémantiques et géométriques, dit modèle de partitions spatiales hiérarchiques, a été proposé dans [Rigaux 94] [Rigaux 95]. Cependant, il ne s'adresse qu'à un type de données particulier (zonages emboîtés).

## 1.4 Contribution de la thèse : Définition d'un processus d'intégration de BDG

L'objectif de cette thèse est de définir un **processus de constitution de BDG multi-représentation** issues de différentes BDG (figure 1). Dans ce but, **un processus d'intégration** de bases de données classiques sera étendu aux bases de données géographiques. Il englobera un processus **d'appariement** afin de relier les instances des différentes bases représentant les mêmes phénomènes du monde réel.

Cette solution paraît le meilleur compromis possible favorisant la **réutilisation** des nombreuses représentations existantes et étendant leur portée par une intégration de celles-ci. Le résultat de cette intégration n'est pas d'obtenir une seule représentation mais de permettre **l'interopérabilité** entre les bases et de relier les instances des différentes représentations représentant le même phénomène du monde réel.

Cette thèse ne pourra pas traiter de l'intégration de tous les types de BDG. Elle se concentrera uniquement sur les bases de données vectorielles planimétriques et n'abordera pas les aspects temporels. De même, cette thèse ne s'attaquera pas l'intégration de BDG réparties (sur un réseau), les représentations des BDG seront migrées sur un seul site (BDG centralisée).

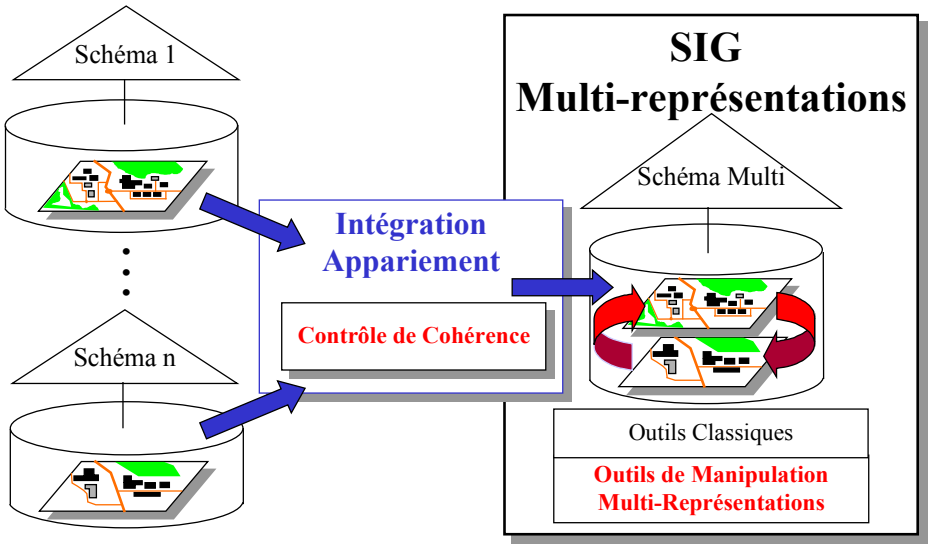



figure 1 : SIG multi-représentation obtenu à partir
d'un processus d'intégration et d'appariement

Cette thèse se situe donc à la frontière de **plusieurs domaines**.

Tout d'abord, elle reprend les travaux déjà réalisés en intégration de BD classiques et sur l'appariement de données géographiques. Elle s'inscrit aussi dans la philosophie actuelle de partage de l'information, de la réutilisation et de l'interopérabilité entre les SGBD géographiques (Open GIS [Buehler et McKee 96], ISO/TC 211 [ISO 96], CEN/TC 287 [CEN/TC 287 96]). Par exemple, des projets comme OpenGIS ont pour but de rendre interopérables des SIG au niveau de leur modèle de données. Pour cela, ils définissent : un ensemble de types basiques de géométries qui peuvent être utilisés quelle que soit l'application ; un ensemble de services qui permettent l'accès à des SIG répartis et le partage de données. L'intégration devrait être la phase suivante de ce genre de projet, car elle permet de générer un schéma intégré et des règles de passage. L'intégration de BD est l'approche la plus puissante pour rendre les BD interopérables. Enfin, ces travaux sont connexes aux problèmes de généralisation de représentations géographiques.

Ces travaux d'intégration et d'appariement de BDG s'insèrent naturellement dans le cadre :

- des recherches du laboratoire COGIT de l'IGN. En effet, ils profitent des compétences en matière de généralisation [Lagrange et Ruas 94] [Ruas et Lagrange 95] et de SIG orienté objet [David et al. 93 c], acquises au COGIT. De plus, ces travaux ont des retombées à la fois pour les recherches engagées par le COGIT et dans une vision à moyen terme pour la production et la gestion des BDG à l'IGN. Effectivement, d'un point de vue applicatif, la constitution de BD multi-représentations facilite la propagation des mises à jour [Kemppainen 94] et le contrôle qualité [Grateau 94]. De plus, la BD multi-représentation peut aussi servir de plate-forme pour tester les outils de généralisation. Enfin, ces travaux sont complémentaires avec les recherches qui viennent d'être engagées sur l'interopérabilité entre les différents SIG au COGIT.
- des recherches au niveau national dans le cadre :

    - de l'axe B1 **Programme de Recherche SIG** (PSIG) du groupe CASSINI, pour les applications géographiques multi-échelles [Dumolard et al. 95] [Raynal et al. 96],
    - du pôle **Bases de Données Spatiales** (BDS) pour les BDG multi-représentations et l'intégration de schémas géographiques [Laurini 95],
- des projets du Comité Européen des Responsables de la Cartographie Officielle (CERCO) pour favoriser l'intégration et de la diffusion des données des agences cartographiques nationales au niveau européen : produit SABE (Seamless Administrative Boundaries of Europe) du groupe **MEGRIN**[2] [MEGRIN 96] [Salgé 95].

## 1.5 Plan de la thèse

La thèse s'articulera autour de deux axes : l'intégration de schémas de BDG et l'appariement des données. Le plan de cette thèse est donc le suivant :

La **partie 2** sera consacrée à un **état de l'art.** Il décrira les nouvelles applications rendues possibles par les BD multi-représentations. Puis, les différents types d'intégration de BD géographiques seront présentés. Enfin, les **techniques d'appariement** des instances des BDG proposées dans la littérature seront exposées.

Dans la **partie 3**, une **approche formelle** de l'intégration de BDG sera exposée. Elle présentera les différentes méthodes d'intégration pour les BD classiques et justifiera le **choix de la méthode d'intégration** déclarative à trois phases de [Spaccapietra et al. 92] pour intégrer les BDG. Ensuite, une **taxonomie des conflits d'intégrations pour les BDG** sera avancée. Pour chaque conflit (différence entre les éléments des BDG à intégrer), si des solutions ont déjà été développées, elles seront décrites et analysées.

Dans la **partie 4**, l'ensemble des conflits à résoudre ayant été analysé et une méthode d'intégration de BD classiques ayant été retenue, les **extensions nécessaires pour les BDG** seront développées. Dans cet objectif, les trois phases de la méthode vont être étendues. Dans la première, la **pré-intégration**, les BDG seront enrichies en rajoutant des données, méta-données et des mécanismes de traductions nécessaires à leur intégration. Dans la seconde, la **déclaration des correspondances,** le langage de déclaration sera étendu, pour tenir compte des spécificités des BDG. Dans la dernière phase, **l'intégration** proprement dite, une technique d'intégration sera adoptée pour chaque déclaration et chaque conflit en fonction de l'objectif à atteindre. Pour regrouper les données représentant le même phénomène du monde réel, le processus d'intégration doit utiliser un processus complexe : l'appariement.

Le processus d'appariement sera décrit dans **la partie 5**. Nous commencerons par exposer le **processus générique d'appariement** servant de squelette aux processus d'appariement spécifiques que nous avons proposé. Celui-ci sera illustré par le **processus d'appariement prototype** développé pour les données de type routier à différentes échelles : les outils utilisés et les résultats obtenus pour les différentes entités seront présentés. L'étape d'appariement permet également d'enrichir les BDG multi-représentations par des opérations multi-représentations, décrivant les différences entre les données. Ces opérations seront décrites ainsi que les méthodes permettant de les extraire.

Pour conclure, dans la **partie 6**, la **contribution** de cette thèse en terme de processus d'intégration et d'appariement de BDG sera synthétisée. Pour chaque application multi-représentation de l'état de l'art, nous montrerons, les avancées résultant de la BD multi-

[2] Multipurpose European Ground-Related Information Network

représentation. Les **extensions** possibles de cette thèse seront aussi évoquées tel que l'adaptation de ce processus pour des BDG réparties sur un réseau.

# 2. Etat de l'art sur les applications multi-représentations et l'intégration de BDG

L'objectif de cette partie est d'avoir une approche générique des applications multi-représentations et des différents niveaux d'intégration de BDG possibles afin de situer cette thèse par rapport aux travaux existants. On décrira donc les applications nécessitant une BDG multi-représentation dans le chapitre 2.1, puis on rappellera les travaux réalisés en terme d'intégration de BDG dans le chapitre 2.2. Ces travaux seront regroupés en fonction du niveau de l'intégration. Finalement, les processus permettant d'intégrer les instances des BDG (l'appariement) seront exposés (2.3).

## 2.1 Nouvelles applications multi-représentations

Sept types d'applications multi-représentations [Devogele 97] sont décrits :

- les trois premiers types répondent aux besoins des utilisateurs : cartographie électronique multi-représentation (2.1.1), analyse multi-représentation (2.1.2) et navigation multi-représentation (2.1.3).
- les quatre suivants sont plus spécifiques aux producteurs de bases de données géographiques : contrôle de cohérence (2.1.4), dérivation de bases de données ayant une représentation hétérogène (2.1.5), propagation des mises à jour (2.1.6) et serveur de données multi-représentation (2.1.7).

### *2.1.1 Cartographie électronique multi-représentation*

Sur une **carte papier**, le volume d'information et l'emprise de la carte sont directement conditionnés par l'échelle qui est fixe. Le nombre de thèmes et leur degré de détail sont donc limités pour produire une carte lisible. Si l'utilisateur veut disposer de plusieurs représentations d'une même zone, il lui faut obligatoirement plusieurs cartes et faire lui-même le rapprochement entre ces cartes.

Ces contraintes sont liées au support papier et doivent être dépassées par les **cartes électroniques**. Ainsi, plusieurs représentations des mêmes données à des échelles différentes et selon différents thèmes, à des époques distinctes doivent pouvoir être affichées. L'échelle et les thèmes sélectionnés par le système ne doivent donc pas être fixés. Le choix se fera en fonction de **critères de sélection cartographique** [Devogele 97]. Ces critères sont :

- les **circonstances** d'utilisation : pour la marine, la classe « feu de navigation » sera sélectionnée uniquement pour la navigation de nuit.
- la **densité** d'information de la zone : un automobiliste utilisera en ville une carte au 1 : 10 000 et une carte au 1 : 100 000 en campagne.
- la **catégorie** de l'usager : l'échelle d'une carte pour un piéton doit être plus grande que celle pour un automobiliste.
- l'**intérêt** de la zone : les militaires [Michel 96] ont besoin d'informations détaillées sur la zone de conflits pour la gestion de la tactique et du terrain et d'informations moins détaillées sur une zone plus large pour la gestion de la logistique et des déplacements.
- la **date** désirée : pour des raisons juridiques, des représentations à des dates différentes doivent être gérées.

- la **distance** entre les objets à visualiser : pour des cartes routières, l'échelle et les thèmes sélectionnés peuvent aussi varier durant l'application.
  Si nous utilisons une représentation autour du point A et que nous voulons nous déplacer vers le point B éloigné (figure 2), la démarche intuitive [Furmas et Bederson 95] est de changer de représentation pour disposer d'une représentation moins détaillée sur laquelle les points A et B apparaissent ensemble à l'écran (fenêtre Marne-la-Vallée). Puis de revenir sur la première représentation autour de B (fenêtre Zoom Intelligent(2)), afin de disposer à nouveau d'une représentation ayant un niveau de détail suffisant. Cette méthode est plus naturelle et plus rapide que de garder la même représentation et de se déplacer pas à pas vers B.

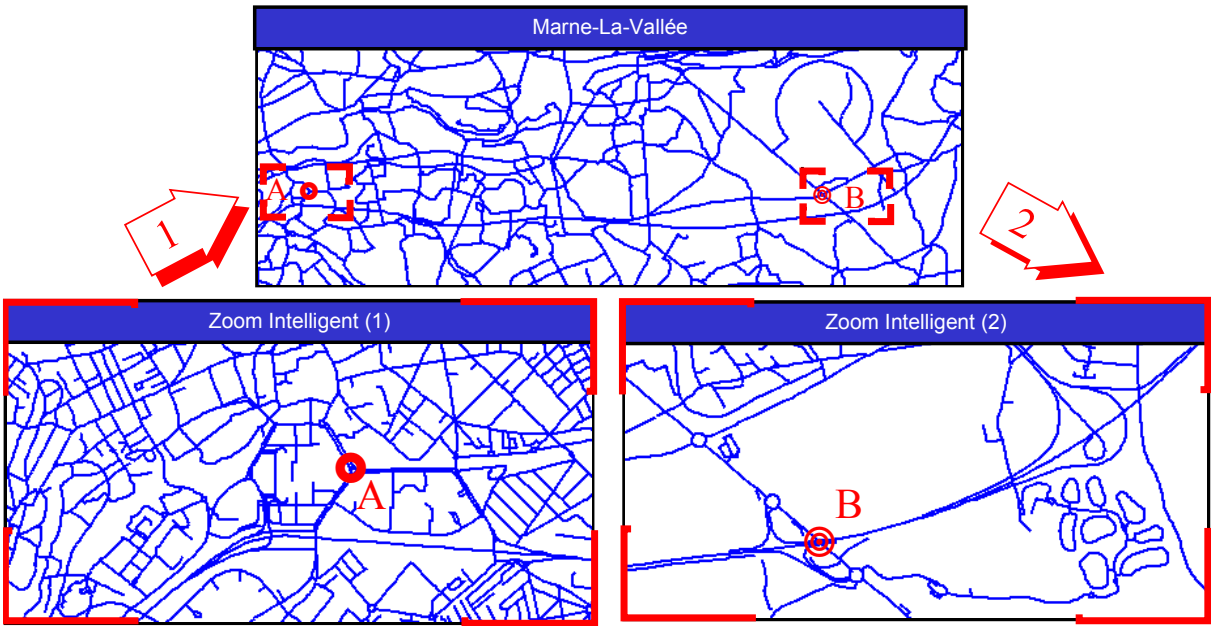



figure 2 : Déplacement vers un point éloigné en utilisant plusieurs représentations

Plusieurs cartes électroniques multi-représentations sont déjà disponibles, mais privilégient seulement certains critères de sélection cartographiques (Annexe 7.2.1).

### *2.1.2 Analyse multi-représentation*

Plusieurs représentations s'avèrent nécessaires pour des traitements plus complexes tels que l'analyse d'impact ou la simulation de phénomènes [Devogele 97]. De plus, ces analyses multi-représentations concernent des domaines variés tels que l'environnement, le transport, l'urbanisme ou encore le géo-marketing. Deux réalisations significatives sont maintenant détaillées : un projet de gestion des feux de forêt et une application de gestion de la flore.

#### 2.1.2.1 Gestion des feux de forêt

Yuan et Albrecht [Yuan et Albrecht 95] ont pu, à partir de l'interview d'experts en gestion de feux de forêt, discerner trois catégories de vues spatio-temporelles nécessaires à un système de gestion des feux de forêt (avec des échelles différentes et des unités de temps différentes) :

- **la vue de gestion des risques de feux** : cette vue permet la prévention des risques des départs de feux de forêt à une petite échelle (entre le 1 : 250 000 et le 1 : 1 000 000) pour une journée donnée en fonction du climat et de son évolution.
- **la vue de gestion des feux en cours** : cette vue est utilisée pour deux types de problèmes : pour le suivi des feux en cours, dans une même forêt, et pour la gestion des sinistres

engendrés par les feux. Elle localise à un instant t (avec une granularité variant entre la seconde et l'heure) les zones de départ de feux, les zones brûlées, et les zones en feux, à une grande échelle (entre le 1 : 5 000 et le 1 : 25 000). Elle permet de gérer aussi l'intensité du feu, le parcours emprunté par les feux et le taux de forêt brûlée.

- **la vue de gestion des anciens feux** : cette vue permet de gérer les traces laissées par les feux sur une longue période (entre la décade et le siècle) à une échelle moyenne (entre le 1 : 25 000 et le 1 : 250 000).

Ces trois vues nécessaires pour un système de gestion des feux de forêt, doivent être reliées pour favoriser **la transmission des informations** recueillies au niveau d'une vue et utilisées par une autre. Par exemple, les nouveaux feux enregistrés dans la vue de gestion des feux en cours doivent être signalés à la vue gestion des anciens feux. De même, les cicatrices laissées par les feux sont utilisées par la vue gestion des risques de feux, une zone brûlée constitue en effet un pare-feu.

#### 2.1.2.2 Gestion de la flore menacée

La forêt primitive de l'est de l'île de Mauï (Hawaii) possède une flore rare qui est menacée par la colonisation de plantes importées du continent, ayant un développement rapide. R. Myers [Myers 97] a donc réalisé une étude multi-échelle sur l'extension des plantes étrangères dans la forêt primitive de l'est de l'île de Mauï. L'utilisation de plusieurs représentations est motivée par le fait que les différentes caractéristiques du paysage intervenant dans l'analyse sont détectables à des échelles différentes :

- **échelle du contexte** : type d'habitat, occupation du sol, routes et autres informations topographiques, sur une zone de 4047 Ha.
- **échelle du sujet** : les différents types de canope (couvert de la forêt) sur une zone de 30 Ha.
- **échelle des détails** : le nombre et la qualité de chaque essence inventoriée et la présence d'animaux sur une zone de 5m x 50m.

Beaucoup d'autres analyses multi-représentations ont été ou sont en cours de réalisation (Annexe 7.2.2).

### *2.1.3 Navigation multi-représentation*

Dans le domaine du transport, pour les applications d'aide à la navigation (calcul d'itinéraire, navigation embarquée,…), une méthode de raisonnement avec différents niveaux de détails, de résolution, de précision, est une bonne démarche [Mark 89] [Timpf et al. 92] [Car et Frank 94] [Langou et Mainguenaud 94].

Par exemple, trois niveaux de détail peuvent être employés :

- le réseau de communication entre les villes principales,
- le réseau routier principal,
- le réseau routier détaillé.

Il faut noter que ces réseaux doivent être reliés de manière hiérarchique : ainsi selon [Langou et Mainguenaud 94], un nœud ou une arête d'un réseau peut être composé d'un ensemble de nœuds et d'arêtes d'un réseau de niveau d'abstraction plus fin.

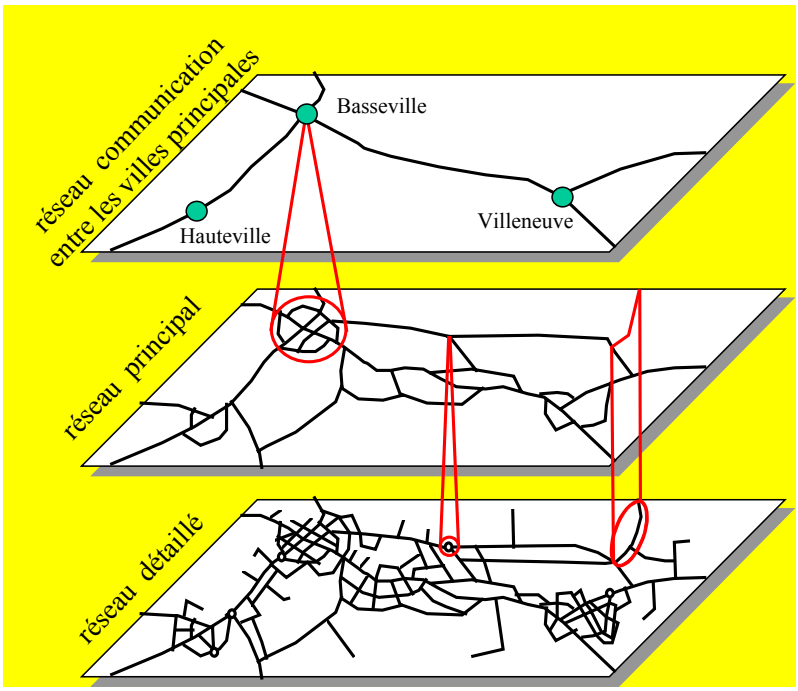



figure 3 : Navigation multi-représentation

### *2.1.4 Contrôle de cohérence*

Le contrôle de cohérence est l'une des classes d'applications requise par les producteurs de BDG. Actuellement, ces derniers ont défini et saisi plusieurs bases de données géographiques. Chaque base répond à un problème spécifique et contient une unique représentation du monde réel. Chaque phénomène du monde réel est représenté une fois et une seule dans la BDG. Ces bases sont appelées **mono-représentations**.

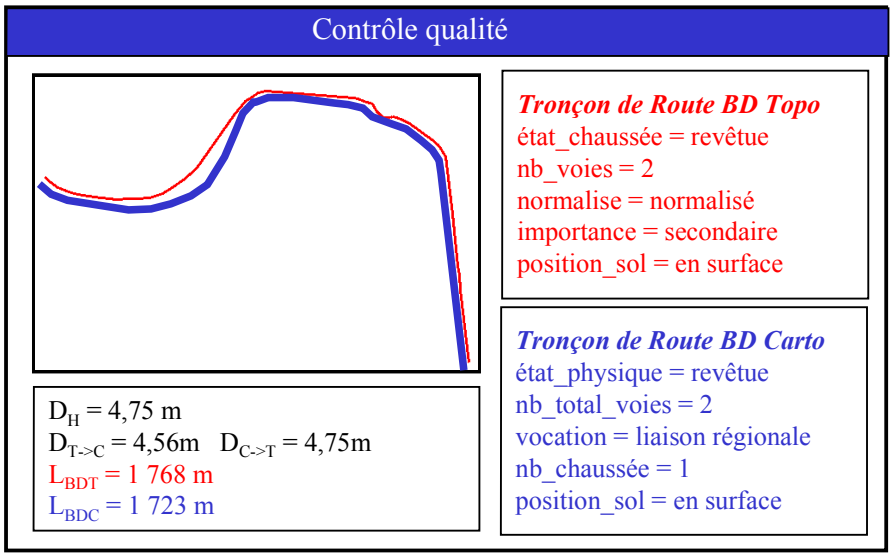



figure 4 : Contrôle qualité

Le regroupement des différentes représentations dans une seule base peut contribuer au **contrôle qualité**[3]. En effet, les incohérences vont pouvoir être détectées et corrigées en comparant les valeurs des instances provenant des différentes bases. Par exemple, dans la figure 4, la sémantique et la géométrie de deux tronçons homologues sont comparées.

Par la suite, la concomitance des différentes représentations va faciliter le **maintien de la cohérence** (assurance qualité[4]).

Une information **qualité** [David 97] (précision géométrique, précision sémantique…) peut être obtenue aisément pour les représentations les moins détaillées par confrontation de ces représentations avec la représentation la plus détaillée. Plusieurs contrôles de cohérence ont déjà été réalisés (annexe 7.2.3).

### *2.1.5 Dérivation de bases de données ayant une représentation hétérogène*

Les BD géographiques multi-représentations permettent de dériver des représentations spécifiques ayant des critères de sélection cartographiques variables suivant la zone. Cette représentation dérivée est dite **hétérogène**.

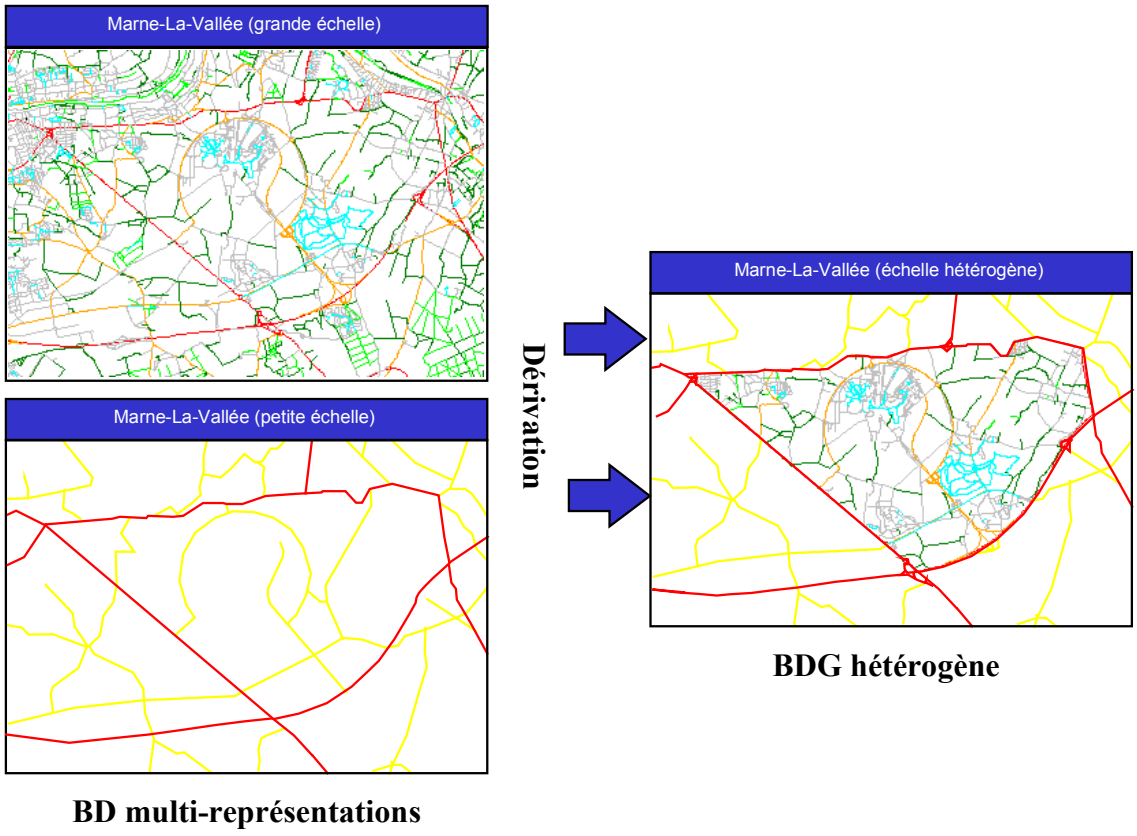



figure 5 : Exemple de base de données ayant une échelle hétérogène, dérivée à partir d'une BDG multi-représentation

[3] Ensemble des actions de mesures, d'examens, d'essais, de calibrage d'une ou plusieurs caractéristiques d'un produit et de comparaisons aux exigences spécifiées en vue d'établir leur conformité [Norme ISO 8402 94].

[4] Ensemble des actions préétablies et systématiques réalisées au fur et à mesure de la production et qui seront nécessaires pour donner la confiance appropriée en ce qu'un produit satisfera aux exigences données relatives à la qualité [Norme ISO 8402 94].

Par exemple, dans la figure 5, une base de données ayant une représentation hétérogène a été dérivée (fenêtre Marne-la-Vallée (échelle hétérogène)), elle est constituée de données détaillées (fenêtre Marne-la-Vallée (grande échelle)) et de données peu détaillées (fenêtre Marne-la-Vallée (petite échelle)).

GEOROUTE [IGN 96 c] est un exemple de BD routières ayant une échelle hétérogène. Elle gère des données détaillées (≈ 1 : 10 000) pour les zones urbaines (agglomérations de plus de 100 000 habitants) et reprend les données de la BD CARTO (≈ 1 : 100 000) pour les autres zones. Cette base a été obtenue en remplaçant les données des zones urbaines la BD CARTO par de nouvelles données détaillées équivalentes.

### *2.1.6 Propagation des mises à jour*

La gestion de plusieurs BDG indépendantes représentant le même espace exige une mise à jour en parallèle de chaque BDG, lors de la prise en compte d'une modification. Ce parallélisme entraîne des risques d'incohérences et un coût de mise à jour global considérable. La présence de toutes les représentations, dans une seule BD multi-représentation permettrait de diminuer les coûts et le temps nécessaire à la mise à jour, de garder les bases cohérentes et donc de mettre à jour plus souvent. Dans cet objectif, il faut définir un processus de **propagation des mises à jour** de la représentation la plus détaillée, qui sera mise à jour manuellement, vers les représentations les moins détaillées, qui seront mises à jour semi-automatiquement [Kilpeläinen 95] [Kemppainen 94] (figure 6).

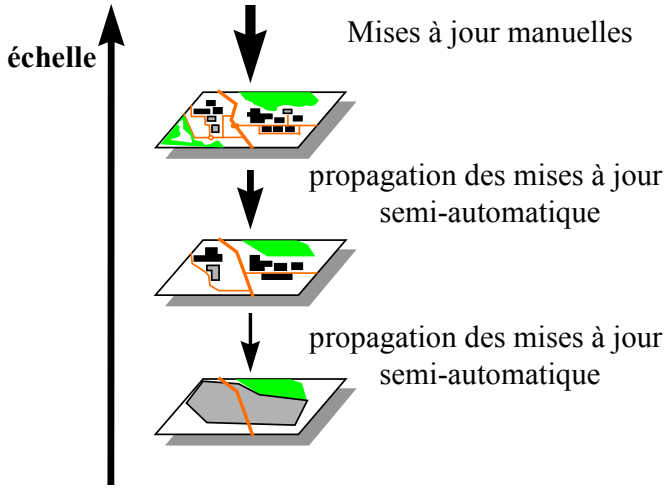



figure 6 : Propagation des mises à jour

Une mise à jour ne peut être propagée sans s'assurer qu'elle ne crée pas de perte d'information ou d'incohérence (superposition d'une route et d'une maison, suppression de relation, …). La propagation des mises à jour demande donc une étude approfondie et fait l'objet d'une nouvelle action de recherche à l'IGN.

### *2.1.7 Serveur de données multi-représentation*

Pour les producteurs de données géographiques, les BD multi-représentations sont les bases les plus appropriées pour servir de **serveur de données** et pour gérer l'ensemble de leurs données [Sleath et Perry 96]. En effet, elles améliorent la qualité des représentations

géographiques, en confrontant les différentes représentations, permettent un maintien de la cohérence, facilitent la mise à jour et autorisent la dérivation de BD hétérogènes.

Cet emploi des BD multi-représentations comme serveur de données, doit dépasser le cadre des producteurs de données géographiques. Effectivement, les SIG s'insèrent de plus en plus au cœur du système global d'information des utilisateurs de données géographiques (collectivité territoriale, gestionnaire de réseau, administration, …) et sont exploités par plusieurs utilisateurs, pour plusieurs applications [U.S. Government 94].

Or, d'après [Boudon 93], les informations géographiques, sont recueillies dans une logique d'autarcie, c'est-à-dire pour les besoins propres de l'institution, voire d'un service. Cette autarcie entraîne des saisies inutiles, une qualité médiocre des données et une mauvaise gestion des mises à jour. Pour éviter ces trois problèmes, Boudon propose de fédérer des informations de différentes provenances pour qu'un grand nombre d'utilisateurs puisse s'en servir. Ce SIG serait constitué à partir des meilleures sources d'information afin d'obtenir les représentations nécessaires pour l'ensemble des utilisateurs. De plus, pour chaque donnée, un unique propriétaire serait chargé d'assurer la fiabilité et de gérer les mises à jour.

### *2.1.8 Conclusion sur les nouvelles applications multi-représentations*

Nous avons donc constaté qu'un grand nombre d'applications nouvelles peuvent tirer profit des **BD multi-représentations**. Effectivement, ces bases autorisent des applications qui sont coûteuses ou impossibles pour des BD mono-représentations. Elles rendent aussi plus naturelle la **visualisation** de données géographiques, en proposant à l'utilisateur, à chaque instant, la représentation la plus appropriée à son application. Les BD multi-représentations permettent de plus une meilleure **manipulation** des données issues des différentes représentations. Ainsi, l'analyse conjointe d'un ensemble de phénomènes corrélés, ce qui peut difficilement être décrit dans la même représentation, et une navigation sans perte d'information entre ces représentations sont rendues possibles.

## 2.2 Les différents niveaux d'intégration des BDG

Actuellement, il existe une multitude de BDG qui possèdent chacune leur représentation[5]. Ces représentations peuvent être regroupées à l'intérieur d'un même SGBD géographiques, pour former une BD multi-représentation. Ainsi, une BD multi-représentation sera obtenue sans avoir recours à la généralisation. La BDG peut être :

- **centralisée sur un seul site**, l'opération qui consiste à regrouper les données sur un seul site est appelée **migration**,
- **répartie** [Öszu et Valduriez 89] sur des sites distants reliés par un réseau.

De plus, l'intégration de BDG peut aussi être employée :

- pour assembler plusieurs BDG ayant des emprises limitrophes ([Laurini 96], [MEGRIN 96]),
- pour réutiliser des données dans un nouveau contexte ([Dangermond 89], [Breunig et Perkhoff 92], [Stephan et al. 93]),
- pour obtenir une BDG commune optimale en terme de qualité et de non redondance ([Gouvernement du Québec 92], [Nyerges 89]),

Nous allons décrire quatre niveaux d'intégration [Devogele 97] des BDG :

- les BDG multi-couches obtenues par regroupement (2.2.1),

[5] Par exemple, 402 sources d'information en France dans le domaine de l'environnement [Morel 96]

- l'intégration des méta-données : les catalogues (2.2.2),
- l'intégration de la sémantique des BDG (2.2.3),
- l'intégration complète de BDG (2.2.4).

### *2.2.1 Les BDG multi-couches*

Dans une BDG multi-couches, les représentations géographiques des différentes BDG sont **regroupées pêle-mêle** dans une seule BDG. De nombreux SIG (Arc/Info®, GeoConcept®,…) proposent ce type de base. Une **couche géométrique** est alors employée pour chaque représentation. Ces couches cohabitent côte-à-côte, sans relations entre elles, chacune avec son schéma et ses données (figure 7).

Pour des problèmes de cartographie électronique multi-échelle, quelques règles peuvent alors contrôler l'emploi de l'une ou de l'autre couche suivant l'échelle d'affichage. Par exemple, en changeant d'échelle graphique, à partir d'une certaine échelle, le système va automatiquement changer de représentation. Cependant, il faudrait inclure dans la définition des règles, d'autres paramètres comme la densité de la zone affichée.

Pour les autres applications multi-représentations, les BDG multi-couches ne sont pas satisfaisantes. Ainsi, la transmission des informations recueillies au niveau d'une représentation, ne peut pas être réalisée lors d'une analyse multi-représentation (les représentations n'étant pas reliées).

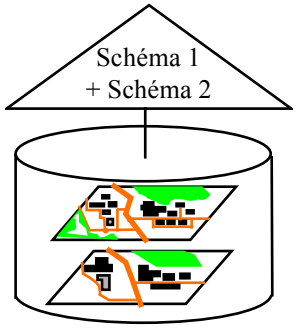



figure 7 : Les systèmes de BDG multi-couches

Pour résoudre ce problème ont été définis :
- des outils permettant de fusionner les objets des différentes couches pour les intégrer,
- des structures reliant les objets.

#### 2.2.1.1 Outils d'intégration des données des BDG multi-couches

L'outil le plus utilisé pour fusionner les géométries est la **superposition de couches** (**Overlay**) [Frank 87] [Dougenik 80] [Pullar 93] [Demirkesen et Schaffrin 96] [Harvey et Vauglin 96]. Cet outil calcule des intersections de deux représentations en employant des tolérances autour des noeuds pour fusionner d'autres noeuds. Pour relier des couches vecteurs et rasters des **outils d'intégration** ont aussi été définis [Piwowar 90]. D'autres outils ont encore été développés comme **l'agrégation de surfaces** [Flowerdew 92] ou le raccordement de géométrique par des **transformations élastiques** (Rubber-sheeting) [Fagan et Soehngen 87].

Malheureusement, tous ces outils souffrent de ne pas savoir si les deux points qui vont être fusionnés représentent le même phénomène du monde réel [Flowerdew 92]. De plus, les

géométries fusionnées à l'aide de ces outils peuvent être erronées [Veregin 89] [Chrisman et Lester 91], [Flowerdew 92].

#### 2.2.1.2 Structures pour relier les objets des représentations

Pour éviter les problèmes liés à la fusion, et pour conserver les données initiales, une autre technique consiste à relier les données par des structures. Ainsi, pour des données de type réseau à différents niveaux de détail [Langou et Mainguenaud 94] a proposé une structure **d'hypergraphe** (un sommet (ou une arête) d'un graphe est composé de sommets et d'arêtes du réseau plus détaillé). D'autres comme [Jones et al. 96] ou [Shepherd 92], se sont tournés vers une structure de type **multimédia** (les objets sont des noeuds composés d'autres noeuds).

Cependant, aucun ne précise comment créer les liens : [Shepherd 92] laisse le processus d'intégration des données qui repose sur un regroupement visuel des informations à la charge de l'utilisateur.

Enfin, les BDG multi-couches souffrent d'un manque de **description globale**. En effet, les utilisateurs doivent savoir dans quelle représentation ils peuvent trouver l'information qu'ils cherchent et si sa qualité est suffisante pour son application. Cette absence de regroupement des méta-données rend la manipulation des données complexe.

### *2.2.2 Intégration des méta-données : les catalogues*

Pour fournir une description globale, des catalogues [Uitermark 96] [Stephan et al. 93] peuvent être renseignés. Ils servent d'interfaces à l'utilisateur pour choisir sa représentation en fonction des données qu'elle contient, de l'emprise de la représentation, de la qualité des données, …

Ces catalogues décrivent aussi bien des représentations d'une BDG centralisée multi-couches que des représentations réparties sur un réseau. Chaque représentation est donc qualifiée par des **méta-données** définies par l'administrateur du catalogue ou répondant à une norme ( [CEN/TC 287 95], [Federal Geographic Data Committee 94]).

Ainsi, le projet américain « **Alexandria Digital Library** » [Frew et al. 95] a créé une bibliothèque électronique répartie sur les informations géographiquement référencées. Cette bibliothèque dispose de différentes cartes électroniques à différentes échelles et d'une interface qui permet de changer de carte.

De même, GEO2DIS [GEO2DIS 97] est un système client-serveur conçu pour consulter via Internet des données de SIG hétérogènes. Pour cela, l'utilisateur va poser une requête sur des méta-données (mode de représentation (vecteur, raster), échelle, date de la dernière mise à jour, …) et obtenir les données.

**Yaser** [Yaser 96] a enrichi le concept de catalogue en reliant les **contextes** (spécifications sémantiques globales de la BDG) des bases. Par exemple, le contexte d'une base sur la gestion des routes sera relié au contexte d'une base sur la gestion des transports en commun.

Pour les représentations réparties sur un réseau, un langage multi-bases qui prend en compte la localisation des données doit être employé. Il peut s'inspirer des **systèmes multi-bases** [Litwin et al. 90] définis pour les BD classiques. Ainsi, les bases réparties sont rendues interopérables. L'utilisateur doit néanmoins connaître la localisation des différentes données et le schéma de chaque base.

Le regroupement des méta-données globales est une première étape nécessaire. Cependant, la sémantique des schémas n'est pas unifiée, et les objets représentant les mêmes phénomènes du monde réel ne sont pas reliés.

### *2.2.3 Intégration de la sémantique des BDG*

Une intégration de la sémantique des schémas est obligatoire dès qu'il s'agit d'uniformiser la description des BDG. Elle consiste à définir une **description unifiée** (le schéma intégré), qui regroupe toute la sémantique des schémas initiaux et des **règles de traduction** qui vont permettre la transformation des données. Pour les BDG, cette intégration s'appuie sur une nomenclature commune ou s'inspire de l'intégration sémantique des schémas des BD classiques ([Jardine et Yazid 89] [Larson et al. 89] [Motro 87] [Spaccapietra et al. 92]) qui sera décrit en 3.1.1.

La **nomenclature** est souvent employée dans le domaine de l'information géographique. Elle permet pour les entités géographiques d'une carte papier ou par extension d'un BDG, de fixer leur nom, leur définition et leurs attributs [Gouvernement du Québec 92]. Cette notion est similaire à la notion de dictionnaire des BD classiques. L'utilisation d'une **nomenclature commune** permet ainsi d'unifier la sémantique.

Par exemple, la BDG SABE (Seamless Administrative Boundaries of Europe) [MEGRIN 96] [Salgé 95] du groupe MEGRIN, contient toutes les unités administratives de 25 pays européens, du niveau pays au niveau commune. Pour concevoir, cette BDG, les nomenclatures ont été intégrées à l'aide d'une nomenclature européenne commune (Nomenclature des Unités Territoriales Statistiques (NUTS) de Eurostat). Ainsi, les niveaux « Ward », « Commune », « Gemeinde » et « Termino Municipal » ont pu être regroupés, car ils sont tous de niveau NUTS 5. Actuellement, un grand nombre de nomenclatures communes (dictionnaire des entités géographiques au gouvernement du Québec [Gagnon et Malboeuf 94], SANDRE [Preux 95], …) ont été définies pour favoriser le transfert des données puis envisager par la suite, l'intégration des BDG [Gouvernement du Québec 92].

D'autres travaux se sont inspirés de l'intégration sémantique des schémas des BD classiques :

- **Nyerges** [Nyerges 89] ont repris les travaux déjà réalisés pour les BD classiques ([Larson et al. 89]) sans vraiment tenir compte des spécificités des BDG.
- **Worboys et Deen** [Worboys et Deen 91] ont aussi repris les mêmes travaux et ont pris en compte les conflits de mode de représentation de la géométrie (3.2.2.4) et les conflits de zonage incompatible (conflits de fragmentation (3.2.3.3)).
- **Breunig** et **Perkhoff** [Breunig et Perkhoff 92] ont proposé une intégration logique à base de vues.
- **Stephan** [Stephan et al. 93] pour intégrer des BD réparties, a proposé de renseigner les données par des méta-données (qualité, type de saisie,…) de définir un format standard, et de créer des jeux de données virtuels, c'est-à-dire d'utiliser des méthodes encapsulées pour représenter les données des BD initiales selon le format commun.

L'intégration de la sémantique des schémas permet d'unifier la description, mais ne permet pas de relier les instances représentant les mêmes phénomènes du monde réel (objets homologues).

### *2.2.4 Intégration complète de BDG*

Les phénomènes du monde réel sont actuellement représentés dans plusieurs BDG, la réutilisation de telles BDG nécessite un processus d'intégration complète pour :

- unifier la sémantique (intégration sémantique) et les méta-données,
- éliminer les parties redondantes et regrouper les parties complémentaires.

Ce **processus d'intégration** [Spaccapietra et al. 92] est nécessaire étant donné le coût d'acquisition des données géographiques (de cette manière des saisies d'information sont évitées) et permet de fédérer l'information provenant de différentes sources. Il consiste

- à prendre en entrée :
    - un ensemble de bases de données (schémas et populations),
- à produire en sortie :
    - une **description unifiée** des schémas initiaux (le schéma intégré),
    - les **règles de traduction** qui vont permettre la migration des données,
    - des **liens** entre les objets des différentes représentations représentant le même phénomène du monde réel.

Pour les BDG, le processus permettant de définir ces liens est appelé **appariement** ou encore conflation.

#### 2.2.4.1 Intégration complète de BDG centralisées

La BDG centralisée issue d'une intégration complète :

- a pour schéma ; le schéma intégré,
- a pour données ; les données initiales *migrées* selon le format du schéma unifié grâce aux règles de traduction(figure 8). Les données représentant le même phénomène du monde réel sont reliées.

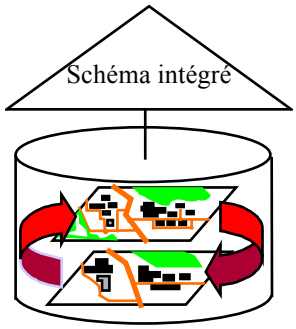



figure 8 : BDG centralisée issue d'une intégration complète

Si les mêmes phénomènes du monde réel sont représentés dans les BDG à intégrer, nous obtenons alors une BDG centralisée multi-représentation. Ainsi, les applications multi-représentation sont rendues possibles, ce qui sera établi en 6.2.2, une fois le processus d'intégration de BDG exposé.

#### 2.2.4.2 Intégration complète des BDG réparties

Pour concevoir une BDG multi-représentation à partir de BDG réparties, un accès global (**le schéma fédéré**), doit être défini à partir des données locales mises à disposition des utilisateurs du système. Ce schéma fédéré regroupe et unifie la sémantique des schémas locaux. Ce type de système est appelé système de **BD fédérées** [Sheth et Larson 90] (figure 9). Un accès global via un schéma fédéré et un accès local via les schémas locaux sont donc

possibles. Le concept de bases de données fédérées fournit ainsi un partage de l'information entre plusieurs SGBD sans remettre en cause l'autonomie et l'intégrité de chaque système de la fédération [Yétongnon et al. 93].

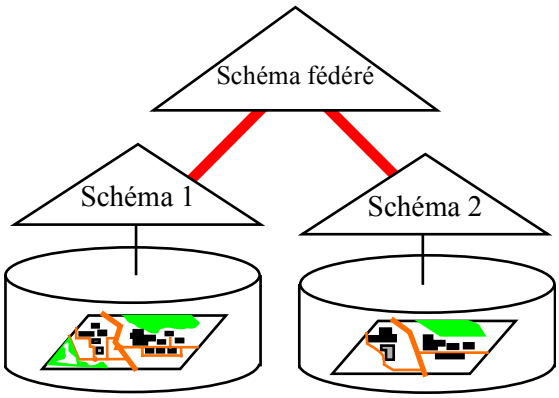



figure 9 : Système de BDG fédérées

Le processus d'intégration est sensiblement le même pour les BD centralisées, à la nuance près que les données ne sont pas migrées dans une BD centralisée. Ces données sont construites à partir de celles des sites distants, lors des requêtes, grâce aux règles de traduction. Il faut alors disposer d'un processus performant d'identification des données homologues.

La gestion de SGBD fédéré est cependant complexe. En effet, l'utilisateur va poser une requête sur le schéma global et le processeur de requêtes réparties va scinder cette requête globale en un ensemble de sous-requêtes. Chaque sous-requête sera exécutée sur un site distant. Il faudra alors récupérer l'ensemble des réponses pour les traduire en une réponse globale. Cela implique une gestion globale des transactions avec l'aide d'un calendrier pour synchroniser et gérer le recouvrement des transactions locales. Qui plus est, les systèmes de gestion des BD réparties peuvent être **hétérogènes**. Il est alors nécessaire de définir des interfaces, pour les faire communiquer.

Les systèmes de BDG fédérées multi-représentation sont une alternative aux systèmes de BDG multi-représentations centralisés, lorsque qu'il n'est pas envisageable de migrer les données des BDG sur un site central. Effectivement, si des traitements importants sur les BD réparties à intégrer existent déjà ou si les données doivent rester réparties pour des raisons de performance, un système centralisé n'est pas adapté.

Dans le domaine des SIG, une seule intégration complète de BDG a été proposée pour définir une fédération entre des BDG juxtaposées [Laurini 96]. Cette intégration reprend la procédure défini dans [Spaccapietra et al. 92] et la surcharge en y ajoutant :

- des pré-traitements afin de résoudre quelques problèmes spécifiques aux BDG,
- un mécanisme de transformation élastique pour le raccordement géométrique aux frontières.

Cette technique n'a pas été étendue aux BDG qui se superposent.

### *2.2.5 Conclusion sur les niveaux d'intégration des BDG*

Quatre niveaux d'intégration des BDG ont été présentés. Ils vont de l'intégration la plus simple (le regroupement pêle-mêle) à l'intégration complète des BDG. La plupart part des travaux réalisés se situent dans les premiers niveaux. Les quelques recherches effectuées dans

les niveaux évolués sont partiels ou ne concernent que des BDG particulières (BDG juxtaposées).

En revanche, le processus d'intégration classique permet d'intégrer complètement les BD initiales. Aussi, le processus d'intégration proposé dans cette thèse s'appuiera sur un processus d'intégration classique afin d'obtenir une BDG intégrée multi-représentation centralisée. Il est donc nécessaire de décrire les outils d'appariements existants qui permettent d'identifier les données géographiques homologues.

Pour conclure, Shepherd [Shepherd 92] résume parfaitement la complexité de l'intégration de BDG :

> *« Il est clair que l'intégration n'est pas produite magiquement par le regroupement d'un flot de données diverses dans un SIG, c'est le résultat d'un effort considérable pour résoudre les contraintes liées aux différentes sources. Mais, avant qu'un SIG puisse être utilisé pour relier les informations diverses, les informations doivent être rendues comparables, compatibles, et cohérentes, ce qui implique un effort humain considérable. »*

## 2.3 Identification des données géographiques homologues : l'appariement

**L'appariement**, encore appelé **conflation**, est le processus consistant à établir les correspondances entre les objets géographiques des différentes bases qui représentent le même phénomène du monde réel. Il est utilisé dans de nombreuses applications manipulant l'information géographique : regroupement de bases de données juxtaposées [Laurini 96], propagation des mises à jour dans une base de données clients [GIS/Trans Ltd 95] [Bucaille 97], recalage de données sur un référentiel [Lupien et Moreland 87] [Lynch et Salford 85], intégration de BDG [Gouvernement du Québec 92], contrôle qualité [Brooker 95], superposition de couches pour fusionner les géométries [Demirkesen et Schaffrin 96] [Schorter et al. 94].

Afin d'apparier les données géographiques, il est nécessaire de développer des outils spécifiques exposés par la suite. En guise de préambule il sera rappelé les techniques d'intégration des données des BD classiques.

### *2.3.1 Identification des données classiques homologues*

Dans le cadre des bases de données classiques, pour identifier les objets représentant le même phénomène du monde réel, les valeurs des identifiants[6] ou attributs clés sont utilisées. Par exemple, pour identifier les instances de deux classes COMMUNES, l'attribut *numéro INSEE* peut être employé. En effet, cet attribut est un identifiant défini par un organisme extérieur, sa valeur sera constante quelle que soit la base. D'autres identifiants de ce type peuvent être utilisés pour d'autres types d'objets (le numéro de téléphone, le code barre,…).

Par contre, les informations géographiques sont recueillies dans une logique d'autarcie. Les identifiants définis sont pour la plupart propres à chaque service. Par conséquent, ils peuvent rarement être utilisés pour identifier des données homologues. Il faut donc intégrer les données géographiques à l'aide d'autres techniques : les techniques d'appariement.

---

[6] Un identifiant est un ensemble d'attributs qui identifie de façon unique un objet dans sa classe.

## *2.3.2 Les mécanismes d'appariement de données géographiques*

### 2.3.2.1 Introduction

A première vue, l'appariement semble être un processus facile à réaliser. En effet, par une visualisation superposée de la géométrie de deux jeux de données, l'humain est capable d'apparier sans aucun problème les objets constituant chacune des bases, de façon quasiment innée, en prenant en compte la **forme** des objets, leur **position** ainsi que leurs **liens** avec les objets les entourant [Lemarié 96].

Par exemple, pour les deux jeux de données superposés de la figure 10, intuitivement, un opérateur va apparier les tronçons 'a' et '1', 'b' et '2', 'd' et '4'. Des appariements plus complexes vont aussi pouvoir être décelés : le tronçon 'c' peut être apparié avec une partie du tronçon '3', le tronçon '5' est apparié avec 'e' et 'f', de même, l'utilisateur détectera visuellement des tronçons non appariés comme 'g'.

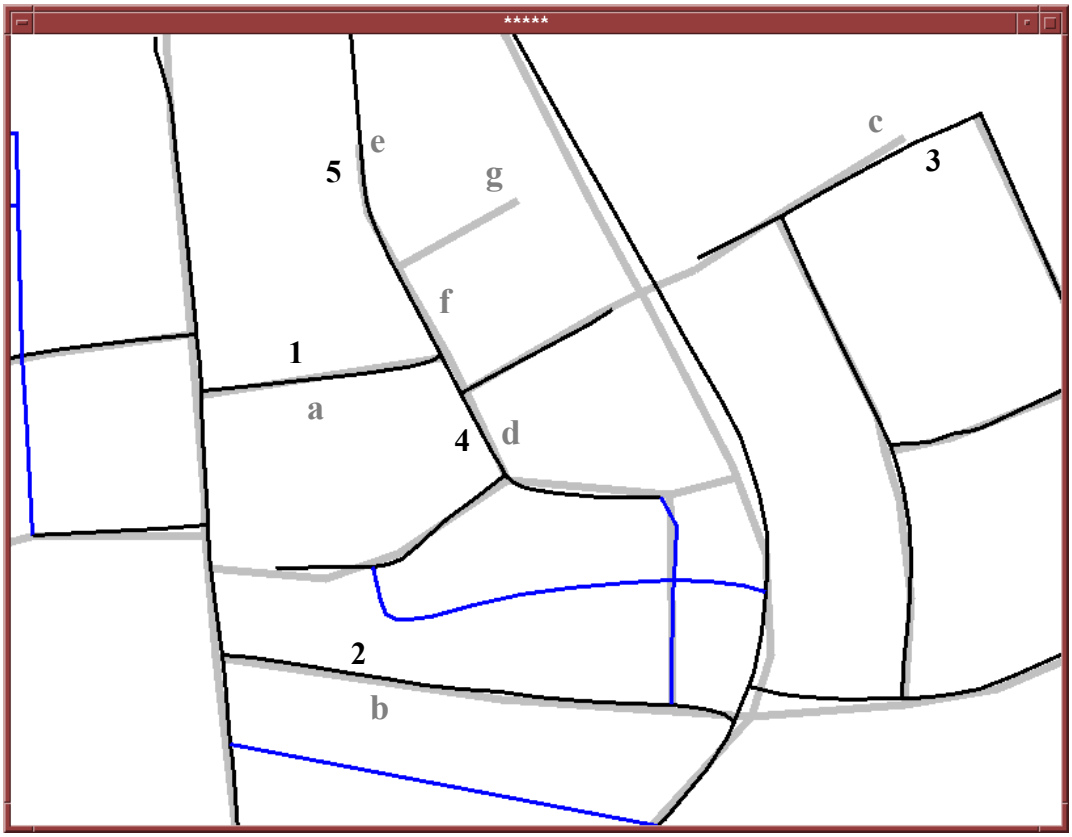



figure 10: Exemple de jeux de données à apparier
( BD TOPO (traits fins) et GEOROUTE (traits épais) )

En effet, un opérateur humain s'appuiera sur la notion intuitive de **ressemblance**. Cette notion est complexe, car elle prend en compte un nombre important d'informations qui ne sont pas forcément cohérentes entre elles.

Pour rendre l'appariement **automatisable**, la difficulté n'est pas dans la définition d'outils à employer pour apparier les données mais dans la transcription de la notion de ressemblance et de notre mécanisme d'appariement visuel en un processus informatique afin de passer outre les conflits. Cette traduction s'avère ardue, compte tenu de la variété des conflits à considérer (3.2). A cause de cette difficulté, certains [Lynch et Salford 85] se sont tournés vers une approche semi-automatique.

Pour distinguer les critères d'appariement, de nombreux outils ont été proposés, chacun étant spécifique à un problème donné.

On distingue trois types d'outils d'appariement :
- l'appariement **sémantique** qui est similaire au mécanisme d'intégration de données classiques (les objets sont alors appariés grâce à la valeur de leur identifiant commun),
- l'appariement **topologique** qui utilise les relations topologiques entre les différents objets pour apparier les données. Si deux relations sont en correspondance, alors cette correspondance doit permettre de trouver les objets homologues unis par cette relation,
- l'appariement **géométrique** qui consiste à apparier les données géographiques par leur localisation et leur forme.

#### 2.3.2.2 Appariement géométrique

Les données à apparier possèdent des géométries de précision différente (et donc des localisations différentes). Par conséquent, l'appariement géométrique doit prendre en compte cette imprécision. Trois types d'appariements géométriques peuvent être adoptés :
1. en définissant une zone d'appariement,
2. en se dotant d'une mesure de distance entre objets,
3. en utilisant d'autres caractéristiques géométriques de la forme de l'objet.

Ces trois types peuvent être utilisés isolément ou conjointement.

##### *2.3.2.2.1 Zone d'appariement*

Le premier type d'appariement géométrique **s'appuie sur l'inclusion des objets géométriques d'une des bases, dans des zones** définies à partir de la géométrie de l'autre base : la zone d'appariement. La base servant à définir les zones est appelée **base de référence**, la seconde, **base à comparer**.

Un objet ne peut être considéré comme apparié avec un objet de la base de référence que s'il appartient à la zone de cet objet. Pour les objets surfaciques, la première zone d'appariement possible est la surface de l'objet. Cette zone d'appariement est dilatée afin de prendre en compte l'imprécision des données.

###### *2.3.2.2.1.1 Rectangle englobant minimum*

Le **rectangle englobant minimum** (figure 11) est défini comme le plus petit rectangle contenant la géométrie d'un objet. Les côtés du rectangle peuvent être orientés parallèlement à l'axe des x et à l'axe des y : on obtient alors le **rectangle englobant minimum x, y**.

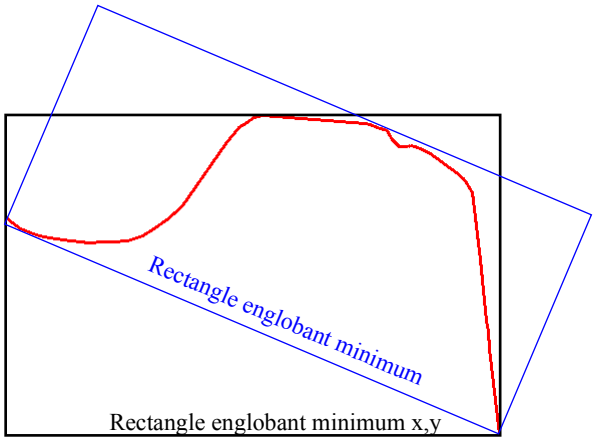



figure 11 : Rectangles englobants minima

Deux objets sont appariés, à l'aide des rectangles, si l'objet de la base à comparer est inclus géométriquement dans le rectangle englobant minimum de l'objet de la base référence. Cette zone d'appariement est généralement dilatée pour tenir compte des imprécisions en bordure du rectangle.

Nous n'évoquerons pas les zones d'appariement de forme différente telles que le cercle circoncrit à la figure. Les rectangles englobants ne tiennent pas compte de la forme de l'objet mais de son emprise globale et pour définir des zones plus précises, les zones tampons peuvent être utilisées.

*2.3.2.2.1.2 Zone tampon (buffer zone)*

Une zone tampon est définie par une géométrie et une distance d. Tous les points dont la distance à un point de la géométrie est inférieure à d appartiennent à cette zone tampon (figure 12). Un objet géométrique de la base à comparer est apparié avec un objet de la base référence s'il est inclus dans sa zone tampon. Cette zone d'appariement est très pratique car elle permet de modéliser assez finement l'imprécision de la géométrie.

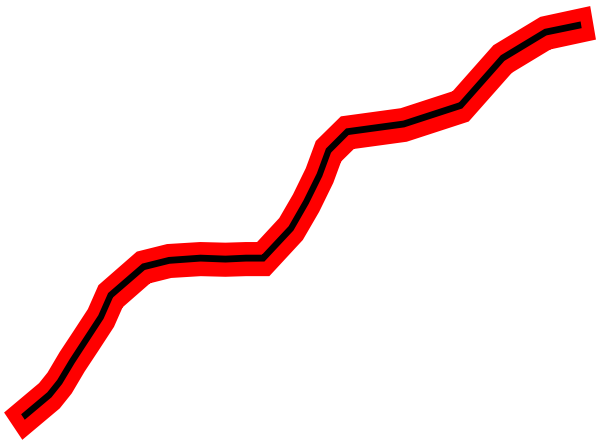


figure 12 : Zone tampon

*2.3.2.2.1.3 Bande Epsilon*

La zone tampon est un cas particulier de la bande Epsilon [Perkal 56]. Cette dernière consiste **à considérer les points et les segments composant les polylignes et à leur assigner une**

**zone de tolérance**. Pour cela on associe à chaque point un cercle de tolérance dont le rayon varie selon la nature du point qu'il représente. Ensuite, les cercles associés à chaque extrémité de segment sont reliés par leurs tangentes communes afin de former la bande de tolérance. Cette technique permet d'apparier les polylignes au niveau de leur composant (segments, points). La figure 13 donne des exemples d'appariement grâce à la bande Epsilon. Cette technique a été utilisée entre autres par [Gabay et Doytsher 94].

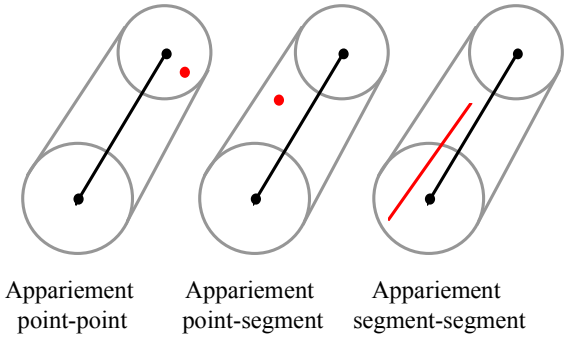



figure 13 : Bande Epsilon

*2.3.2.2.1.4 Diagramme de Voronoï*

Le Diagramme de Voronoï [Gold 90] [Djadri 96] est un pavage de l'espace à partir de sites (point, segment) et à l'aide d'une distance. Le critère est : « un point de l'espace appartient à la zone d'un site si et seulement si sa distance à ce site est le minimum de ses distances à tous les sites ». Ce pavage de l'espace peut être utilisé pour définir des zones d'appariement pour les objets de la base de référence. En effet, un pavage de l'espace peut être obtenu pour chaque objet de la base en fusionnant les pavés de ses sites. Par exemple, pour la figure 14, un pavage de l'espace a été obtenu pour des habitations à partir du diagramme de Voronoï sur l'ensemble des sites de ces habitations. Ainsi, une habitation de la base à comparer sera appariée à une habitation de la base de référence si elle est incluse dans son pavé. Cette technique n'a pas encore été utilisée dans le cadre de l'appariement.

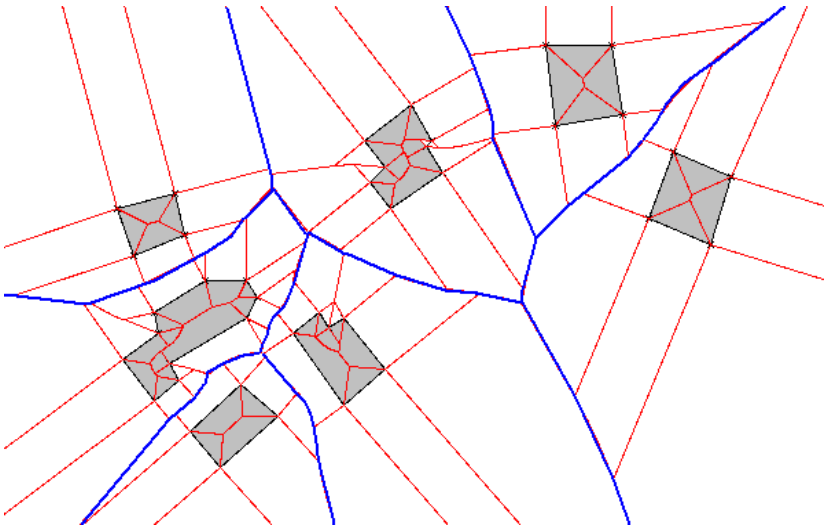


figure 14 : Exemple de pavage issu du diagramme de Voronoï, pour des habitations

*2.3.2.2.1.5 Conclusion sur les zones d'appariement*

La méthode de zones d'appariement présente un certain nombre d'avantages : elle permet l'appariement entre un objet de la base de référence et plusieurs objets de la base à comparer. Elle est donc particulièrement adaptée pour des bases à des échelles différentes. De plus, elle permet de tenir compte de l'imprécision des données.

Par contre, ce type de méthodes souffre de leur caractère binaire. Elles ne tiennent pas compte de situations particulières comme les objets qui intersectent une zone d'appariement. Pour résoudre ce problème, un critère d'appariement supplémentaire et un seuil d'appariement selon ce critère peuvent être utilisés.

Par exemple, pour apparier les habitations du Cadastre avec les habitations de la BD TOPO, [Lemarié 96] a utilisé comme zone d'appariement la surface des habitations du Cadastre et comme critère d'application supplémentaire : le rapport (aire d'intersection / aire habitation du Cadastre) et un seuil de 60 %.

L'inconvénient majeur de ces méthodes réside dans la détermination des bons paramètres. Si la détermination d'un paramètre, s'appuie le plus souvent sur des tests visuels, cette valeur n'est pas forcément adaptée à toute l'emprise de la surface [Branly 97]. Enfin, ces méthodes ne sont pas symétriques, le résultat obtenu en posant la base 1 en référence n'est pas identique au résultat obtenu en posant la base 2 en référence.

***2.3.2.2.2 Distances entre objets***

Pour définir la ressemblance entre deux objets, les distances entre ces objets semblent être des mesures pertinentes. Deux objets de classe correspondante sont appariés si la distance sélectionnée est inférieure à un seuil.

Pour apparier deux objets ponctuels, la **distance Euclidienne** s'impose.

Pour apparier deux objets linéaires, trois distances vont être décrites (distance moyenne, distance de Hausdorff, distance de Fréchet).

Pour apparier deux objets surfaciques, une distance a été proposée.

*2.3.2.2.2.1 Distance moyenne*

Pour qualifier la généralisation d'une ligne avec la ligne d'origine, McMaster [McMaster 86] a proposé d'utiliser la mesure de la surface de déplacement totale divisée par la longueur de la ligne du jeu de données de référence. Cette mesure est simple. En effet, le contour de la surface de déplacement est défini à partir des deux lignes à apparier, du segment reliant les noeuds initiaux et du segment reliant les noeuds finaux. Cette surface est divisée par la longueur de la ligne d'origine pour rendre cette mesure indépendante de la longueur. Elle permet de rendre compte du déplacement moyen dû à la généralisation. Par exemple, pour les deux lignes de la figure 15, la surface de déplacement totale est représentée en gris, si la ligne de référence est la ligne pointillée. La mesure obtenue est la surface en gris divisée par la longueur de la ligne de référence.

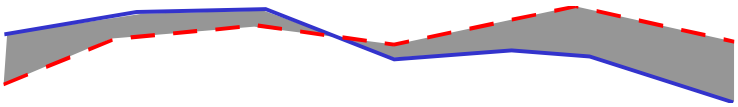


figure 15: Surface de déplacement totale / longueur de l'arc original

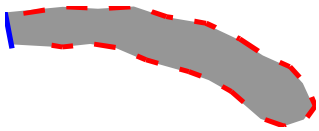


figure 16 : Distance moyenne faible produisant un appariement erroné

Cette mesure peut être rendue symétrique en divisant la surface par la moyenne des longueurs des arcs. Ainsi on obtient la **distance moyenne** entre les deux lignes. Cette distance n'a pas été utilisée dans le cadre de l'appariement.

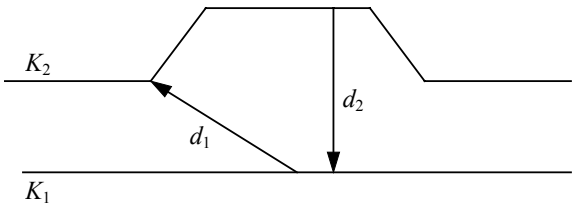
Cette distance moyenne doit toujours être couplée avec une distance maximum (distance de Hausdorff ou de Fréchet décrit ci-dessous), sinon elle risque de donner des résultats surprenants. Par exemple, pour la figure 16, la distance moyenne calculé entre ces deux lignes sera faible, tandis qu'elles ne s'apparient pas visuellement.

#### *2.3.2.2.2.2 Distance de Hausdorff*

La première distance possible pour rendre compte de l'écart maximum entre les lignes ($K_1$, $K_2$) est la distance de Hausdorff [Hausdorff 19] [Hangouët 95].

La distance de Hausdorff (figure 17) est la plus grande des deux **composantes** :

- $d_1$ qui est la plus grande valeur de la fonction « distance » de $K_1$ à $K_2$,
- $d_2$ qui est la plus grande valeur de la fonction « distance » de $K_2$ à $K_1$.




$$d_1 = \max_{p_1 \in K_1}\left[\min_{p_2 \in K_2}\left[dist(p_1, p_2)\right]\right]$$

$$d_2 = \max_{p_2 \in K_2}\left[\min_{p_1 \in K_1}\left[dist(p_2, p_1)\right]\right]$$

$$d_H = \max(d_1, d_2)$$

figure 17 : Exemple et définition de la distance de Hausdorff.

Cette distance est utilisée dans le cadre de l'appariement par le laboratoire COGIT. Un algorithme permettant d'apparier les polylignes à partir des composants de la distance de Hausdorff et d'un seuil, a été développé sous $GéO_2$ [Stricher 93], [Raynal et Stricher 94].

#### *2.3.2.2.2.3 Distance de Fréchet*

La distance de Fréchet s'appuie sur la propriété suivante : toute polyligne orientée est équivalente à une application continue f :[a, b] → V ou a, b ∈ ℜ, a < b et V est l'espace vectoriel. La distance de Fréchet ($d_F$) [Fréchet 06] est la suivante :

Soient f :[a, a'] → V et g :[b, b] → V' deux polylignes et || || la norme usuelle,

$$d_F(f, g) = \inf_{\substack{\alpha:[0,1]\to[a,a'] \\ \beta:[0,1]\to[b,b']}} \max_{t\in[0,1]} \left\| f(\alpha(t)) - g(\beta(t)) \right\|$$

Une illustration intuitive de la distance de Fréchet est la suivante : un maître et son chien suivent deux chemins. Ils avancent ou s'arrêtent à volonté, indépendamment l'un de l'autre, mais il ne peuvent pas revenir sur leurs pas. La distance de Fréchet entre ces deux chemins est

la longueur minimale de la laisse qui permet de réaliser une progression de concert satisfaisant ces conditions.

Par rapport, à la distance de Hausdorff, la distance de Fréchet a l'avantage de calculer la distance uniquement sur des couples de points qui auraient pu être mis en correspondance visuellement, ce qui n'est pas le cas pour la distance de Hausdorff. Par exemple, pour la figure 17, le point de $K_2$ servant à calculer $d_1$, n'est pas le correspondant intuitif du point de $K_1$, c'est simplement le point le plus proche. La distance de Fréchet est donc plus proche de la notion de distance maximum entre deux lignes. Cette distance n'a pas encore été utilisée dans le cadre de l'appariement du fait de sa complexité. Le calcul pratique de cette distance n'est pas évident, un algorithme de calcul complexe (algorithme d'ordre $O(pq \log^2 pq)$ avec p et q le nombre de segments des polylignes) est donné dans [Alt et al. 92] [Alt et Gadau 95] et un algorithme simple (O(pq)) donnant une approximation discrète de la distance dans [Eiter et Mannila 94].

*2.3.2.2.2.4 Distance entre surfaces*

Peu de distances et de mesures entre surfaces ont été proposées pour apparier deux surfaces. [Vauglin 97] propose la **distance surfacique** suivante :

soit A et B deux surfaces, $D_S(A,B) = 1 - \frac{S(A \cap B)}{S(A \cup B)}$ *(ou S(X) est la valeur de la surface de X)*

Cette distance permet de mesurer le rapport des surfaces communes par rapport à l'union des surfaces. Elle vaut 1 si A et B se superposent et 0 si elles sont disjointes.

Cette distance surfacique a été généralisée pour des appariements 1:n [Bel Hadj Ali 97] et donne la mesure suivante :

$$D_S(A_i, B) = 1 - \frac{\sum_{i=1}^{n} S(A_i \cap B)}{S(\bigcup_{i=1}^{n} A_i \cup B)}$$

D'autres mesures ont aussi été proposées comme la **probabilité d'association** [Phalakarn 91] [Lemarié 96] :

$$P_a(A,B) = \frac{S(A \cap B)}{S(A)}$$

ou la **fonction de ressemblance** [Bel Hadj Ali 97] :

$$F_r(A,B) = \frac{S(A \cap B)}{\min(S(A), S(B))}$$

Ces mesures sont trop récentes pour pouvoir donner un avis critique.

*2.3.2.2.2.5 Conclusion sur l'utilisation de distance*

Les distances sont les mesures les plus fiables pour décrire les différences de positions. Par contre, l'utilisation de distances pose des problèmes. En effet, la quasi totalité de ces distances sont utilisables uniquement pour apparier un objet à un autre objet de même dimension.

Pour apparier un objet à un ensemble d'objets (appariement 1-n) ou pour apparier un ensemble d'objets à un ensemble d'objets (appariement n-m), les distances doivent être utilisées avec précaution. Par exemple, pour la figure 18, « a » doit être apparié avec « 1 » et

« 2 ». Or si nous utilisons des distances entre « a » et « 1 » et entre « a » et « 2 », les distances vont donner des résultats contraires à l'appariement. En effet, les distances maxima seront portées par les extrémités. Pour résoudre ce type de problème, une solution consiste à définir des mesures non symétriques des lignes les plus petites vers les lignes les plus grandes [Stricher 93].

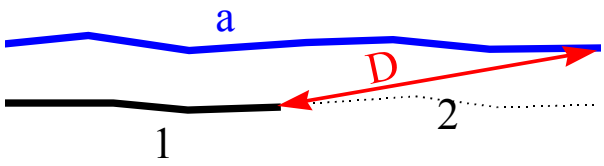



figure 18 : Distance entre deux lignes de longueur différente

De même, pour apparier des objets **de dimension différente**, des opérations transformant la géométrie de l'objet en un objet de dimension égale à la dimension de l'objet à apparier peuvent être utilisés. En général ces opérations diminuent la dimension de la géométrie de l'objet :

- la **squelétisation** qui transforme une surface en une ligne,
- le **calcul du barycentre** qui transforme une surface ou une ligne en un point.

Des opérations augmentant la dimension de la géométrie de l'objet peuvent également être envisagées. Par exemple, pour transformer une ligne en surface, la géométrie peut être **boudinée**, un tronçon de route ayant un attribut largeur égale à 3m, sa géométrie linéaire peut être transformée en géométrie surfacique.

#### *2.3.2.2.3 Ressemblance de forme*

La ressemblance des formes ne s'arrête pas à la distance entre deux géométries, d'autres critères ont été utilisés ou proposés. Par exemple [Jones et al. 96] proposent d'utiliser les mesures définies par McMaster [McMaster 86] pour apparier les objets linéaires. De même, Kidner [Kidner 96] a soumis un ensemble de mesures pour les objets surfaciques.

Pour les **objets linéaires,** les critères proposés pour apparier en fonction de la forme [McMaster 86] [Mustière 95] sont :

- le rapport des longueurs des arcs,
- le rapport des longueurs moyennes des virages,
- la différence entre les directions de chaque segment,
- le rapport du nombre de points intermédiaires,
- le rapport des sommes des angles entre les segments (les angles sont en valeur absolue et calculés entre 0 et π (figure 19)),
- le rapport du nombre de virages (un virage étant défini comme le sous-arc compris entre deux points d'inflexion).

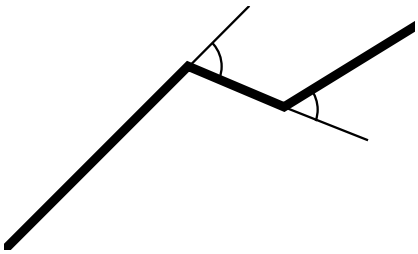


figure 19 : Angles entres les segments

Des mesures plus complexes ont aussi été proposées dans [Plazanet 96].

Pour les **objets surfaciques** les critères proposés pour apparier en fonction de la forme [Kidner 96] sont :

- le rapport des aires,
- le rapport des périmètres,
- le rapport des compacités (périmètre/aire),

- le rapport des élongations, défini par (1,27 × aire / longueur) ou la longueur est l'axe principale de la surface,
- le rapport des excentricités (axe principal/ axe secondaire),
- le rapport des allongements {longueur/largeur),
- le rapport des rectangularités (aire du polygone/aire du rectangle englobant minimum),
- le rapport des distances maximum au centroïde,
- le rapport des distances radiales au centroïde (distance du centroïde à la frontière, cette distance est calculée sur le cercle trigonométrique selon un intervalle fixé),
- la comparaison des transformations de Fourrier,
- la comparaison des dimensions fractales.

Kidner a testé ces différents descripteurs pour des surfaces simples représentant le même phénomène du monde réel à différentes échelles. Sa conclusion était la suivante : l'aire, la rectangularité et l'excentricité sont de bons critères. La distance radiale est intéressante. Par contre, le périmètre est un très mauvais critère et la transformation de Fourrier ainsi que la dimension fractale ne sont pas pertinentes.
Pour les **surfaces complexes** (surface ayant des trous ou étant formée de plusieurs composants surfaciques) d'autres caractéristiques d'appariement peuvent être ajoutées. On peut citer, le nombre de trous, le nombre de composants connexes, l'enveloppe convexe et le nombre d'Euler.

Il existe ainsi un grand nombre de critères pour apparier les objets selon leur forme. Malheureusement, ces critères ne peuvent être utilisés que pour des appariements 1-1 et ce type d'appariement n'est pas suffisant. Il faut bien entendu ajouter au moins un critère de localisation, sinon deux objets ayant la même forme à des kilomètres de distance pourraient s'apparier.

De plus, la sélection des critères pertinents est complexe, elle est fonction des caractères communs aux données à apparier.

#### *2.3.2.2.4 Conclusion sur l'appariement géométrique*

Trois types de méthodes d'appariement géométrique ont donc été exposés (les zones d'appariement, les distances, les ressemblances de forme).

Un quatrième type de méthodes d'appariement géométrique a été développé : les **méthodes probabilistes** [Servigne 93] [Servigne 94] [Salmeron et Milgram 86] [Phalakarn 91] [Jamet et Phalakarn 89] [Le Men et Jamet 90]. Elles consistent à calculer la **probabilité d'association** pour chaque objet candidat à l'appariement, en fonction de critères (recouvrement, angle relatif, …). Ces méthodes ont été surtout utilisées pour apparier des segments images avec un jeu de données vecteurs. Elles permettent de regrouper plusieurs critères d'appariement dans le calcul de la probabilité.

Les techniques d'appariement géométrique regroupent donc un grand nombre de méthodes possibles. Pour l'utilisateur, il est très difficile de choisir la ou les méthodes à utiliser. Qui plus est, une fois la méthode choisie, il lui reste à fixer les valeurs des paramètres. Les techniques d'appariement géométrique sont donc puissantes mais complexes à utiliser.

### 2.3.2.3 Appariement topologique

L'appariement topologique n'utilise plus la géométrie des objets mais les relations entre ces objets et entre leurs géométries. Il est appelé appariement topologique car il utilise

principalement les relations topologiques. Cependant, d'autres types de relations usuelles peuvent aussi être utilisés comme par exemple, les relations de composition.

L'appariement topologique ne peut pas être employé sans aucun des deux autres appariements. Il est soit utilisé en complément soit à la suite d'un autre appariement.

Par exemple, pour apparier les limites de communes linéaires (figure 20), on peut apparier les communes sémantiquement en utilisant le numéro INSEE. Puis, grâce aux relations topologiques entre les arcs et les faces, les objets représentant la limite entre les communes A et B peuvent être appariés en isolant l'arc ou les arcs reliés par les relations « Face Droite », « Face Gauche » à la géométrie de A et de B.

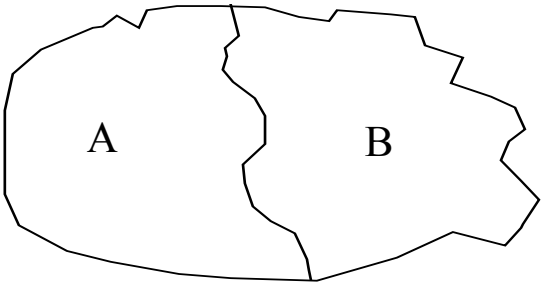



figure 20 : Appariement des limites de communes

Cet exemple a démontré l'utilité de l'appariement topologique dans le cadre des relations entre les surfaces et leurs frontières. Les relations topologiques entre les noeuds et les segments, peuvent aussi être utilisées [Gabay et Doytsher 94] [Phalakarn 91]. Par exemple, une fois un couple d'objets appariés, une prolongation en suivant les relations topologiques permet de diminuer le nombre de recherches d'appariement. Cela est particulièrement vrai pour l'appariement de deux réseaux.

Les relations de **composition** entre le composé et ses composants peuvent aussi être utilisées pour limiter le nombre d'objets appariables aux niveaux des composants, une fois les composés appariés.

### *2.3.3 Conclusion sur l'appariement*

Trois types d'appariement (sémantique, géométrique et topologique) ont été décrits dans ce chapitre. Les techniques d'appariement sémantiques sont déjà communément utilisées dans le cadre des bases de données classiques. Pour les BDG, la littérature a surtout proposé des techniques d'appariement géométrique parfois couplées avec des techniques topologiques. Cet état de l'art sur l'appariement montre que cette tâche est complexe à résoudre, dans le choix des méthodes à utiliser et dans le choix des valeurs de leurs paramètres. De plus, nous pouvons constater que ces trois types sont nécessaires pour apparier des BDG dès que les données présentent des différences importantes. Effectivement, aucun type d'appariement n'est suffisant par lui même.

Hélas, malgré le grand nombre d'applications possibles pour les méthodes d'appariement, les SIG du commerce offrent très peu de fonctions pour apparier les objets sémantiquement équivalents mais qui présentent des différences (localisation, incohérence au niveau des attributs) [Jones et al. 96]. Seules quelques techniques de zone d'appariement sont régulièrement proposées comme le rectangle englobant minimum x, y et la zone tampon [GIS/Trans Ltd 95].

## 2.4 Conclusion sur l'état de l'art

Cet état de l'art a permis d'identifier les nouvelles applications destinées aux utilisateurs et aux producteurs de BDG, qui nécessitent des BDG multi-représentations.

Pour répondre, à ces dernières, nous nous sommes tournés vers l'intégration de BDG. Nous avons donc analysé et regroupé en fonction du niveau d'intégration les différents travaux d'intégration déjà réalisés.

Nous avons alors constaté qu'aucun travail d'intégration n'était suffisant pour intégrer les schémas et les instances des BDG. Nous proposerons donc un processus d'intégration qui s'appuie sur un processus d'intégration de BD classiques.

De même, pour obtenir une intégration complète, notre processus d'intégration inclura un processus d'appariement permettant d'identifier les instances homologues. Cette tâche (l'appariement) est complexe pour les BDG. Un grand nombre de techniques d'appariement ont ou peuvent être utilisées : elles s'appuient sur des données différentes (sémantique, géométrique, topologique) et sont complémentaires.

# 3. Approche formelle de l'intégration de BDG

L'objectif de cette partie est de présenter les méthodes d'intégration classiques et de justifier le choix de celle qui sera étendue aux BDG (3.1). Puis, afin d'expliquer la complexité de la phase d'intégration que nous allons aborder dans cette thèse, on détaillera l'ensemble des conflits à résoudre (3.2), en se focalisant sur les conflits spécifiques aux BDG. Pour chaque conflit, il sera fait mention des solutions proposées.

## 3.1 Les méthodes d'intégration de bases de données classiques

L'intégration de BD consiste à **produire une description unifiée des BD d'origine sans redondance**. Définissant de la sorte un système global, les bases initiales deviennent donc **interopérables**, c'est un problème émergeant à l'heure actuelle, car les applications doivent réutiliser, re-exploiter des données mémorisées dans des bases existantes mais indépendantes [Parent et Spaccapietra 96].

Dans ce chapitre, nous présenterons les deux types d'intégration (l'intégration structurelle et l'intégration sémantique) en 3.1.1 puis les deux méthodes d'intégration sémantique : l'intégration procédurale (3.1.2) et l'intégration déclarative (3.1.3). Une méthode déclarative [Spaccapietra et al. 92] sera retenue pour être étendue aux BDG. Cette dernière sera présentée et son choix sera justifié.

### *3.1.1 Intégration structurelle ou sémantique*

Deux types d'intégration se distinguent : l'intégration sémantique et l'intégration structurelle. **L'intégration structurelle** [Thieme et Siebes 93] [Geller et al. 92] a pour but de réduire les redondances au niveau des attributs et des méthodes. Les classes ayant des propriétés en commun sont donc unifiées soit en les fusionnant soit en les intégrant dans un graphe d'héritage. Par exemple, Thieme et Siebes proposent d'intégrer les classes suivantes :

| | |
|---|---|
| classe **Carré** : n-uplet ( x : entier<br>y : entier<br>largeur : entier) | classe **Rectangle** : n-uplet ( x : entier<br>y : entier<br>largeur : entier<br>longueur : entier) |

Ils utilisent alors le concept d'héritage qui permet d'obtenir le schéma intégré suivant :

| | |
|---|---|
| classe **Carré** : n-uplet ( x : entier<br>y : entier<br>largeur : entier) | classe **Rectangle** hérite de **Carré** : n-uplet (<br>longueur : entier) |

L'intégration structurelle permet d'obtenir un schéma optimisé en terme de redondance mais ne se soucie pas de la cohérence sémantique du résultat (les rectangles ne sont pas des carrés). Ce type d'intégration peut se justifier si les concepts d'héritage[7] et de sous-typage[8] sont distincts. Or, la plupart du temps ces deux concepts sont fusionnés.

L'**intégration sémantique** [Mannino et Effelsberg 84] [Batini et al. 86] [Motro 87] [Larson et al. 89] [Jardine et Yazid 89] [Spaccapietra et al. 92]**,** par contre, est une approche tenant

[7] l'héritage définit la réutilisation : si U hérite de T alors la structure et le comportement de T sont réutilisés pour U

[8] le sous-typage définit la substitution autorisée entre les types: si le type U est un sous-type de T alors les instances de U sont des instances de T

compte de la signification des classes (leur sémantique) et de l'information contenue dans les classes (leur population). Cette correspondance peut être :

- en **extension** : les instances des classes en correspondance représentent les mêmes phénomènes du monde réel. Par exemple, les objets de la classe ROUTE de la BD TOPO PARIS et les objets de la classe ROUTE de GEOROUTE PARIS représentent les mêmes phénomènes du monde réel.
- en **intension** : les classes sont des abstractions équivalentes de phénomènes du monde réel. Par exemple, la définition de la classe ROUTE de la BD TOPO PARIS correspondent à la définition la classe ROUTE de la BD TOPO MARSEILLE.

Pour les correspondances en extension, les relations de correspondance entre les classes peuvent être de différents types. La relation **d'équivalence** est la plus utilisée. Deux classes sont en relation d'équivalence si pour chaque objet de la première classe, il existe un objet dans la deuxième classe représentant le même phénomène du monde réel et vice versa. D'autres relations ensemblistes peuvent aussi être employées comme *l'inclusion, l'intersection,* ou la *disjonction*. L'intégration sémantique doit donc permettre d'obtenir un schéma unifié sans redondance, ne violant pas la sémantique des bases d'origine.

Dans la littérature, deux types de méthodes sont proposés : les méthodes procédurales (3.1.2) et les méthodes déclaratives (3.1.3).

### *3.1.2 Méthodes procédurales*

Les méthodes procédurales [Motro 87] [Batini et al. 86] s'appuient sur un ensemble d'opérations de restructuration prédéfinies. Par exemple, Motro [Motro 87], pour son prototype Superviews définit un ensemble d'opérations permettant de **mettre en conformité** les classes à intégrer et de les **relier**. Prenons les deux classes : EMPLOYE et ÉTUDIANT (figure 21 a), elles ont un attribut commun *nom* et un attribut propre *diplôme* pour ÉTUDIANT et *salaire* pour EMPLOYE. Les quatre opérations d'intégration proposées par Motro sont les suivantes :

- **Meet A and B into C** (figure 21 b), qui définit une **classe mère** pour les deux classes, construite à partir des attributs communs. Dans notre exemple, Meet ÉTUDIANT and EMPLOYE into PERSONNE produit les classes suivantes :
    - PERSONNE : n-uplet (nom),
    - ÉTUDIANT : n-uplet (diplôme) hérite de PERSONNE,
    - EMPLOYE : n-uplet (salaire) hérite de PERSONNE.
- **Join A and B into C** (figure 21 c) qui crée une **classe fille** pour les deux classes, construite en unifiant les attributs. Join EMPLOYE and ÉTUDIANT into ETUDIANT-SALARIE, produit les 3 classes suivantes :
    - ETUDIANT-SALARIE qui hérite d'ÉTUDIANT et d'EMPLOYÉ,
    - EMPLOYE qui ne change pas.
    - ÉTUDIANT qui ne change pas.
- **Combine A and B into C** (figure 21 d), qui **fusionne** deux classes en une troisième, ces deux classes sont détruites. Les instances de A et B deviennent alors des instances de C. Par exemple, Combine EMPLOYE and ÉTUDIANT into PERSONNE, crée :
    - une classe PERSONNE décrite par trois attributs, *nom* qui est obligatoire et *salaire* et *diplôme* qui sont optionnels.
- **Connect A to B** (figure 21 e), qui permet de relier A et B par une **relation d'héritage**, A devient une classe fille de B. Par exemple, Connect EMPLOYE to PERSONNE crée une relation d'héritage entre ces deux classes.

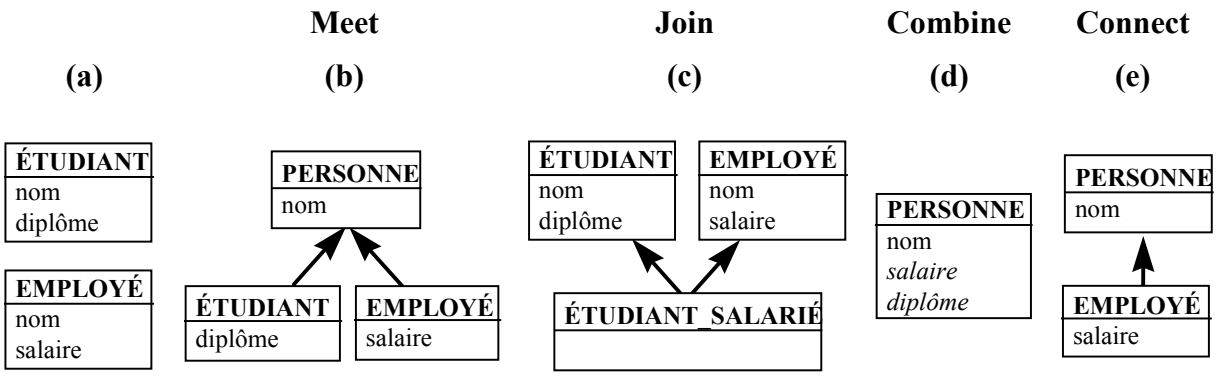



figure 21 : Opérations d'intégration

Cet ensemble d'opérations est complété par un ensemble d'opérations de **mise en conformité** des classes à intégrer qui permettent de faire évoluer les schémas initiaux vers le schéma intégré. Ces opérations sont :

- **Rename** A to B qui remplace le nom de la classe A par B,
- **Aggregate** ($T_1$,…, $T_n$) of A into T qui forme un attribut de type n-uplet qui a pour éléments $T_1$,…, $T_n$,
- **Telescope** T into S qui réalise l'opération inverse de Aggregate,
- **Add** qui crée une classe ou un attribut en affectant des valeurs par défaut,
- **Delete** qui supprime une classe ou un attribut,
- **Fold** A into B, qui inclut une classe fille dans la classe mère.

Cette approche répond au problème d'intégration sémantique (conception d'une structure unifiée) en manipulant directement les schémas à intégrer. Mais elle fait l'hypothèse que l'intégrateur a identifié toutes les correspondances entre les éléments initiaux par ses propres moyens et est à même de concevoir la nouvelle structure en ayant fait tous les choix d'intégration (join plutôt que combine, …). Or, cette étape d'identification des correspondances peut être difficile et délicate. L'intégrateur doit donc avoir une compétence certaine et l'intégration ne doit générer qu'un nombre limité de conflits, ce qui est rarement le cas comme nous l'exposerons dans le chapitre 3.2. Ces deux conditions limitent donc la portée de ce type de méthode.

### *3.1.3 Les méthodes déclaratives*

Les méthodes déclaratives [Mannino et Effelsberg 84] [Jardine et Yazid 89] [Spaccapietra et al. 92], distinguent les deux problèmes d'intégration :

- l'identification des éléments des bases qui représentent les mêmes phénomènes du monde réel (c'est l'étape de déclaration des correspondances),
- la conception d'une structure précise permettant de représenter les instances de l'ensemble des bases d'origine (c'est l'étape d'intégration). Cette seconde étape peut être simultanée ou différée.

Cette séparation facilite l'intégration des BD. Dans le domaine des BDG, les bases ont été constituées pour répondre à des besoins spécifiques très différents d'une base à l'autre. En conséquence, leur intégration est complexe et nécessite une méthode déclarative, qui s'appuie sur un langage précis.

### 3.1.3.1 Déclarations des correspondances

La **déclaration des correspondances** consiste à identifier les éléments homologues (classes, instances, attributs,…) des différentes bases. Elle se focalise sur le phénomène du monde réel représenté par l'élément de la base, sans tenir compte de sa représentation. Deux éléments (classes, attributs, objets,…) sont donc en correspondance s'ils décrivent les mêmes éléments du monde réel (phénomène, ensemble de phénomènes, propriété des phénomènes, …). Ces déclarations sont exprimées indépendamment de la façon dont les schémas seront ensuite intégrés.

Il existe plusieurs méthodes de déclarations de correspondances **entre classes** ( [Spaccapietra et al. 92] [Mannino et Effelsberg 84] [Jardine et Yazid 89]. Nous avons retenu la déclaration des correspondances proposée dans [Spaccapietra et al. 92] [Parent et Spaccapietra 96] pour intégrer les BDG car :

- elle s'appuie sur les extensions permettant d'établir des comparaisons au niveau des phénomènes du monde réel.
- elle propose un formalisme précis, uniforme et complet, autorisant la déclaration des correspondances entre les BD et de leurs différences.

Les autres méthodes n'ont pas été retenues car elles sont complexes (double qualification des correspondances pour [Mannino et Effelsberg 84]) ou inadaptée (intégration de vues externes pour [Jardine et Yazid 89]). De même, Laurini [Laurini 96] avait retenu cette méthode pour intégrer les BDG.

Le formalisme du langage de déclaration des correspondances de [Spaccapietra et al. 92] et [Parent et Spaccapietra 96] sera décrit dans la section 3.1.4.2.

### 3.1.3.2 Intégration des BD

Une fois les relations de correspondances déclarées, les bases doivent être intégrées. Les opérations utilisées sont semblables aux opérations présentées pour les méthodes procédurales. Pour déterminer les opérations sélectionnées, il est nécessaire de se fixer une **stratégie** en fonction de l'objectif de l'intégration.

Dans certains cas, les utilisateurs vont chercher à obtenir un schéma simple avec un minimum de type d'objets de façon à rendre aisée l'utilisation du schéma intégré. Dans d'autres cas, toutes les classes des schémas initiaux devront être conservées dans le schéma intégré, afin de permettre aux utilisateurs de retrouver facilement les données provenant de leur base. Il existe plusieurs **critères** pour définir l'objectif de l'intégration en fonction du type de schéma intégré désiré. [Dupont 95 b] cite six critères :

- la **liberté d'application**, qui définit pour chaque opération si elle peut être appliquée quelle que soit la relation de correspondance (équivalence, inclusion, …),
- la **conservation**, indique si une opération conserve toutes les informations initiales,
- la **précision**, qui signale si l'opération conserve ou dégrade la précision initiale,
- la **complétude**, qui indique si toutes les redondances ont été éliminées par l'opération,
- la **réversibilité**, qui indique si l'opération permet de réaffecter les informations du schéma intégré sur les schémas initiaux.
- l'**unification**, qui indique si la technique crée des éléments qui regroupent toutes les occurrences.

Chaque opération d'intégration répond ou ne répond pas à chacun de ces critères par construction. Des critères plus subjectifs ont été aussi définis comme la **simplicité** du résultat ou **l'optimisation**. Dans le processus de [Parent et Spaccapietra 96], les opérations n'étant pas fixées, une stratégie d'intégration peut être définie et suivie.

### *3.1.4 Présentation du processus classique retenu*

Pour concevoir le processus d'intégration de BDG, nous avons donc retenu comme fondement le processus défini par Spaccapietra et Parent [Parent et Spaccapietra 96] [Spaccapietra et al. 92]. Il se décompose en trois phases : la pré-intégration, l'identification des correspondances et l'intégration (figure 22).

#### 3.1.4.1 La pré-intégration du processus classique

La pré-intégration inclut toutes les activités préliminaires qui ont pour objectif de faire converger les descriptions initiales. Elle consiste à réarranger les schémas en entrée pour les rendre plus homogènes sur les plans sémantique et syntaxique et pour parvenir au même niveau de compréhension des données. Trois étapes la composent :

- le choix du **modèle de données** pour le schéma intégré (le modèle commun). Un mécanisme de traduction des modèles initiaux vers le modèle commun et vice versa doit aussi être déterminé.
- **l'enrichissement** des schémas initiaux. L'acquisition d'informations supplémentaires auprès de l'administrateur est nécessaire, pour interpréter les schémas sans ambiguïté et obtenir la sémantique du monde réel qui ne peut pas être déduite à partir des schémas. En effet, une traduction entièrement automatique des schémas dans le modèle commun est difficilement envisageable du fait de la différence entre la richesse de description des modèles.
- Des **règles de normalisation** sont imposées aux schémas initiaux pour les rendre plus uniformes et diminuer les différences de modélisation.

Cette phase est le plus souvent passée sous silence dans les autres processus d'intégration, l'homogénéité des bases à intégrer étant posée en tant que prérequis, ce qui n'a pas lieu d'être étant donné le développement indépendant des bases à intégrer [Parent et Spaccapietra 96].

#### 3.1.4.2 L'identification des correspondances du processus classique

L'objectif de cette seconde phase est d'identifier et de fournir toutes les correspondances entre les schémas de données au niveau de la sémantique et de l'instanciation. Elle s'appuie sur **la déclaration d'Assertions de Correspondance Interschémas (ACI)** qui mentionne les éléments en correspondance. Elle répond, par-là même à la question suivante :
Pour chaque phénomène du monde réel, quel ensemble d'éléments de la BD1 le représente et quel ensemble d'éléments de la BD2 le représente ?
Puis elle relie ces deux ensembles.

Ces ensembles peuvent être désignés par leurs noms ou spécifiés par une requête (sélection, jointure, union, …). Les **relations ensemblistes ( <RE> )** sont l'équivalence (≡), l'inclusion (⊆), l'inclusion stricte (⊂), l'intersection (∩), la contenance stricte (⊃), la contenance (⊇) et la disjonction (≠).

| **BD1** | **BD2** |
|---|---|
| classe **COMMUNE** : n-uplet ( | classe **COMMUNE_IDF** : n-uplet ( |
| Num_INSEE = entier, | numéro_INSEE = entier |
| Département = string, | département = string, |
| Région = string, | maire = Personne, |
| Population = entier, | géométrie = surface) |
| Géométrie = surface) | |

tableau 1 : Exemple de classes à intégrer

Pour illustrer la déclaration des ACI, l'assertion de l'exemple du tableau 1 va être détaillée. Elle décrit la correspondance entre la classe COMMUNE de la BD1 qui gère l'ensemble des communes de France et la classe COMMUNE_IDF de la BD2 qui décrit l'ensemble des communes de la région Ile de France. Pour cet exemple, on obtient donc l'ACI suivante :

**ACI** : SELECTION(Région = «Ile de France») BD1. COMMUNE ≡ BD2. COMMUNE_IDF

Cette déclaration de correspondance doit être enrichie par une clause spécifiant comment les instances correspondantes sont identifiées dans leur extension. Cette clause est appelée **« Avec Identifiants Correspondants »** (**AIC**). Elle a pour objectif d'indiquer les attributs qui permettent de détecter les couples d'instances représentant les mêmes phénomènes. Dans la BD intégrée, ces instances seront regroupées afin de disposer de toute l'information disponible sur un même phénomène du monde réel. Dans notre exemple, la clause AIC est :

**AIC** : BD1.Num_INSEE = BD2.numero_INSEE

Au sein de l'information disponible pour un même phénomène, une partie de l'information est redondante. Afin d'éviter cette duplication d'information dans la BD intégrée, une clause **« Avec Attributs Correspondants » (AAC)** est utilisée dans les déclarations. Pour notre exemple, l'attribut Département et département décrivent une information similaire, nous obtenons donc la clause :

**AAC** : BD1.Département = BD2.département

Le format général d'une ACI est le suivant :

**ACI** BD1.élément1 <*RE*> BD2.élément2
**AIC** identifiants communs
**AAC** attributs communs

Les ACI forment un langage de déclaration des correspondances, simple et précis. Ce langage doit aussi inclure la **déclaration des conflits** (différences entre les BDG à intégrer).

### 3.1.4.3 L'intégration du processus classique

La troisième phase, l'intégration proprement dite, traite toutes les ACI déclarées dans la phase précédente. Elle répond à trois objectifs :

- Elle doit **résoudre les conflits** décrits dans les ACI. Pour chaque conflit déclaré, une méthode de mise en conformité est choisie. La méthode utilisée dépend de la stratégie d'intégration définie par l'administrateur. Par exemple, l'attribut Num_INSEE sera renommé Numéro_INSEE.
- Elle doit fournir une description intégrée : le **schéma de la base intégrée.** Ce schéma est obtenu à partir des schémas initiaux normalisés, des ACI et de la stratégie d'intégration.

Pour chaque relation de correspondance, une technique d'intégration crée des éléments unifiés dans le schéma intégré. Pour notre exemple, une relation d'héritage entre la classe COMMUNE_IDF et COMMUNE sera créée si la technique de sous-classe (7.3.1.6) est retenue.

- Enfin, l'intégration doit produire les **règles de traduction** des schémas initiaux vers le schéma intégré, et vice versa.

Il est nécessaire d'élaborer une **stratégie d'intégration** afin de sélectionner les opérations adaptées (sous-classe, …) parmi l'ensemble des techniques d'intégration possibles pour chaque ACI et de fixer l'objectif de l'intégration. L'objectif est défini par un ensemble de critères (simplicité, complétude, précision, réversibilité, …) recherchés pour le schéma intégré. Si la phase de déclaration des correspondances est suffisamment riche et la stratégie d'intégration précise, cette phase doit pouvoir être largement automatisée.

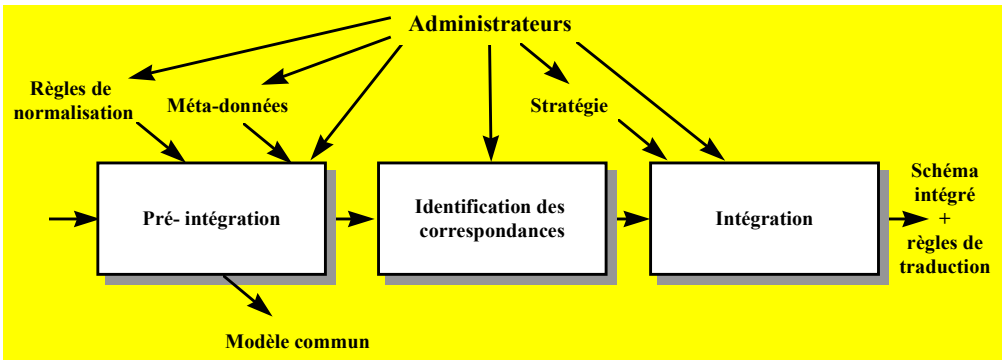



figure 22 : Le processus global d'intégration

### *3.1.5 Conclusion sur les méthodes d'intégration*

Cette description des méthodes d'intégration de bases de données classiques, n'est pas exhaustive, mais seulement suffisante pour aborder la suite des travaux et comprendre la complexité de cette tâche. Des synthèses plus complètes ont déjà été réalisées, entre autres dans [Batini et al. 86], [Dupont 95 b] et [Parent et Spaccapietra 96].

Par contre, il a permis de choisir le processus d'intégration classique [Spaccapietra et al. 92] pour les BDG et de justifier ce choix. Pour les schémas, cette intégration se fera à un niveau conceptuel en utilisant les concepts définis dans UML (Unified Modeling Language) (annexe 7.1).

Ce processus doit cependant être étendu pour tenir compte des conflits d'intégration propres aux BDG qui vont être décrits maintenant.

## 3.2 Taxonomie des conflits d'intégration de BDG

Quant il faut intégrer des classes ayant le même type et la même population, l'intégration est triviale : la classe intégrée est identique aux classes d'entrées. Malheureusement, les phénomènes du monde réel ont été modélisés indépendamment dans les différentes bases de données [Goodchild 90]. Il n'y a donc aucune raison pour que les différents concepteurs aient utilisé la même modélisation du monde réel et de ses phénomènes. De ce fait, les classes à intégrer présenteront certainement des différences dans leur structure ou dans leur population. Ces différences sont appelées **conflits d'intégration**. Il faut pouvoir les identifier afin de les signaler lors de la déclaration des correspondances. Dans cet objectif, nous avons proposé une taxonomie des différents conflits d'intégration.

Pour les bases de données classiques, plusieurs taxonomies des conflits ont déjà été proposées, une des plus détaillées est [Sheth et Kashyap 93] qui contient une trentaine de conflits. La littérature est importante dans ce domaine. [Batini et al. 86], [Comyn-Wattiau 90], [Dayal et Hwang 84], [Dupont95], [Ouksel et Naiman 93] et [Spaccapietra et Parent 91] [Parent et Spaccapietra 96] donnent chacun leur classification.

Dans le cadre de l'intégration de BDG, nous nous restreindrons à une **taxonomie** regroupant les conflits les plus répandus pour les BDG. Le vocabulaire employé est défini en annexe 7.1. Il en résulte **6 classes** :

1. les conflits liés aux sources de données employées pour constituer les BDG (3.2.1),
2. les conflits portant sur les modèles et les méta-données (conflits d'hétérogénéité) (3.2.2),
3. les conflits portant sur la définition des classes, des relations et sur leur instanciation (3.2.3),
4. les conflits liés à la structure utilisée pour représenter les éléments (classe, relation ou attribut) (3.2.4),
5. les conflits portant sur la description des éléments (3.2.5),
6. les conflits de données (3.2.6).

Cette taxonomie permet de décrire les conflits spécifiques aux BDG et de replacer les conflits classiques dans le cadre de l'intégration des BDG. Les solutions proposées dans la littérature, pour chaque type de conflit énoncé, seront exposées.

### *3.2.1 Conflits de sources de données*

Pour saisir les données des BDG, des **relevés** sur le terrain sont pratiqués grâce à des instruments de mesures de distances et d'angles, instruments complétés récemment par les **GPS** (Global Positioning System). Ces processus sont longs (ils nécessitent le déplacement personnes sur le terrain) et de fait coûteux. Pour réduire les coûts et les durées d'acquisition, d'autres sources d'informations sont utilisées comme les **photographies aériennes**, les **images satellitaires** ou les **images radars**. Les BDG peuvent aussi utiliser les **cartes papier** comme source de données. Ces différentes sources de données sont décrites, par exemple, dans [Kavouras et al. 95]. De plus pour constituer des bases complexes, plusieurs sources d'information sont employées.

Il existe donc un **conflit de sources de données** entre deux bases à intégrer quand les types de sources sont différents ou quand les caractéristiques des sources sont différentes [Shepherd 92]. Par exemple la BD TOPO est constituée à partir de photo aériennes et de relevés terrain, alors que la BD CARTO est réalisée à partir de cartes papier au 1 : 50 000.

Les conflits de sources de données engendrent de nombreux conflits : pour la gestion de la troisième dimension (3.2.2.3.2), pour la détermination de la résolution et de la précision (3.2.2.5) et pour l'intégration des données (3.2.5). Les conflits de sources de données seront traités lors de la résolution des conflits « avals » qu'il entraîne. Cependant, il est nécessaire de compléter la BDG par des méta-données qui spécifient la ou les sources de données.

### *3.2.2 Conflits d'hétérogénéité*

Une fois les sources de données connues, il faut se préoccuper des différences qui portent sur l'ensemble des BDG. Pour les BD classiques, ces différences sont regroupées sous le terme de conflit d'hétérogénéité et décrivent essentiellement les conflits de modèle.

Les éléments des bases de données classiques sont, le plus souvent, des abstractions simples de phénomènes du monde réel : les phénomènes sont décrits à l'aide d'attributs de type simple

(entier, chaîne de caractères,…). Par contre, pour les classes d'objets géographiques, la description de la localisation, de la forme et du positionnement relatif des objets est plus difficile. Il existe donc énormément de conflits entre les bases de données géographiques [Parent et al. 96] [Spaccapietra et al. 96] liés à la modélisation du positionnement des éléments (3.2.2.2), à la modélisation de leur altitude (3.2.2.3), au mode de représentation de la géométrie (3.2.2.4), aux méta-données liées aux géométries de la BDG (3.2.2.5) et aux relations topologiques entre objets géographiques (3.2.2.6). Les conflits d'hétérogénéité seront donc étendus à ces conflits spécifiques des modèles de données géographiques.

#### 3.2.2.1 Conflits de modèle

Un modèle **relationnel** « pur » n'est pas adapté à la représentation de l'information géographique [Scholl et al. 96]. En effet, il n'est pas possible d'y définir des nouveaux types de données tel que point, ligne, polygone. Il faut donc utiliser les types de base (entier, réel, …) pour représenter les types géographiques. Par exemple, chaque polygone peut être représenté par une relation stockant l'ensemble de ses sommets. Pour obtenir l'ensemble des points d'un polygone, il faut donc consulter la table décrivant la relation polygone - points et la table des points. Cette opération est lourde à gérer, et oblige l'utilisateur final à maîtriser des niveaux très bas, comme les tables de points, pour poser des requêtes, ce qui est inconcevable.

Comme les modèles standards ne sont pas adaptés, les éditeurs de SIG ont développé leur propre modèle de données géographiques. Les modèles des bases de données géographiques sont donc plus hétérogènes que les modèles des bases de données classiques. On distingue quatre grandes familles [Rouet 91] :

- Les modèles qui gèrent des **primitives graphiques** offrant la possibilité d'associer à chaque primitive un **label**. Ces logiciels sont le plus souvent issue de la CAO (AutoCAD, Microstation [Querzola et Billout 95]).
- Les modèles des **logiciels graphiques qui gèrent leur propre base de données** qui sont des extensions des modèles précédant (Apic Space, GEO Concept [Querzola et Billout 95]).
- Les modèles des **logiciels hybrides**. Ces logiciels conçoivent de façon séparée une base de données géométriques et une base descriptive. Deux modèles de données cohabitent alors : le modèle géométrique et le modèle relationnel (Arc/Info, MGE [Querzola et Billout 95]).
- Les modèles **relationnels étendus** ou **objets**. Les données géographiques y sont représentées comme n'importe quelles données (SDE, Gothic + Lamps [Querzola et Billout 95], Géo2 [David et al. 93 c]).

Il y a ainsi un **conflit de modèle** entre deux BDG à intégrer, quand les modèles de données utilisés sont différents [Laurini 96].

Une **solution** pour résoudre ce conflit consiste à traduire les modèles des bases à intégrer en un **modèle commun** [Shepherd 92]. Ce modèle commun s'appuie le plus souvent, sur des modèles ou langages standards (ou en cours de standardisation) : SQL pour le relationnel, ODMG [Atwood et al. 93] pour l'orienté objet, ISO/TC 211 [ISO 96], CEN/TC 287 [CEN/TC 287 96] Open Geodata Model [Open GIS Consortium 96]. Ils définissent ainsi des types élémentaires pour décrire l'information géographique. Cependant, ce ne sont que des modèles cibles : le problème de conversion entre les modèles d'origine et le modèle commun subsiste et doit être résolu.

#### 3.2.2.2 Conflits de type de positionnement

Pour les BDG, les données doivent être **localisées** sur la Terre. La forme de la Terre peut être déterminée par des mesures astronomiques et dynamiques (pesanteur) mais reste très irrégulière et ne permet pas d'effectuer des calculs aisés. Elle est donc approximée par un ellipsoïde (ellipsoïde d'Hayford, UGI, Clarke,…). La position d'un point sur un ellipsoïde est donnée par la **latitude** et la **longitude**.

Un autre système de coordonnées, **cartésien** défini par trois axes orthogonaux ayant pour origine le centre de la Terre peut être choisi. La position d'un point est alors donnée par les distances X, Y, Z depuis l'origine. Ce type de positionnement est fourni par les récepteurs GPS.

Néanmoins, l'ellipsoïde ou le système tridimensionnel ne sont pas aisés à représenter sur un plan. Une projection des données sur le **plan** est donc réalisée pour faire correspondre les points de l'ellipsoïde aux points du plan. La position d'un point est alors donnée par un couple de coordonnées planaires (X, Y) définissant la distance selon les axes de projection à un point fixe, origine de la projection. Il existe plusieurs **systèmes de projection** (Gauss, UTM, Lambert, …). Chacun de ces systèmes de projection, va engendrer sur le plan, des altérations différentes (déformations des longueurs, des angles ou des surfaces).

Pour des zones de petite dimension (Cadastre d'un ville, …), les données du plan peuvent être seulement rattachées à un réseau de points de référence pour lequel les déformations dues à la forme de la Terre sont négligées.

Le dernier système de géo-référencement est celui des **positionnements indirects** (adresses postales, points kilométriques, …).

Il existe donc cinq grands types de positionnement : position sur un ellipsoïde (longitude, latitude), position sur un ellipsoïde (X,Y,Z)**,** position sur un plan rattaché à un système de projection, position sur un plan et position indirecte (tableau 2 : Exemple de système de positionnement).

| **Position sur un ellipsoïde (latitude, longitude)** | **Position sur un ellipsoïde (X, Y, Z)** | **Position sur un plan rattaché à un système de référence** | **Position sur un plan rattaché à un réseau de points** | **Position indirecte (Adresse postale)** |
|---|---|---|---|---|
| latitude 48º50'47''N<br>longitude 2º24'55''E | X : 4 201 809 m<br>Y : 177 230 m<br>Z : 4 779 462 m | x : 606 450m<br>y : 1 127 090 m<br>Lambert I | x : 50 012m<br>y : 24 233m<br>$X_0$ : 556 438 m<br>$Y_0$ : 1 102 857 m | 2 Avenue Pasteur<br>94160 S$^t$ Mandé<br>France |

tableau 2 : Exemple de système de positionnement

Il y a donc un **conflit de positionnement** [Shepherd 92] [Laurini 96] quand :

- le type de positionnement est différent,
- l'ellipsoïde de référence est différente,
- le système de projection est différent pour deux positionnements sur un plan rattaché à un système de référence,
- le point de référence est différent pour deux positionnements sur un plan,
- le type de phénomène permettant le positionnement indirect est différent. Par exemple, 4 Avenue des Canadiens, Joinville le Pont et N 4 6$^{ième}$ kilomètre.

Pour **résoudre** ces conflits des fonctions de transformation des coordonnées, des mécanismes de recalage [Fagan et Soehngen 87] [Rouet 91] sont utilisés. Ils permettent de transférer les

données d'un système à un autre, mais ils peuvent provoquer des altérations sur les angles et les distances dues aux propriétés des modèles de positionnement.

#### 3.2.2.3 Conflits de gestion de la troisième dimension : la hauteur

La plupart des SIG manipulent des donnés en **2D**, ce qui consiste à représenter tout point seulement à l'aide de coordonnées planimétriques. Cependant la troisième dimension apparaît comme essentielle dans beaucoup de domaines d'application des SIG (géologie, archéologie, architecture, …).

Deux types de conflits peuvent émerger : un conflit de modélisation et un conflit d'abstraction.

##### *3.2.2.3.1 Conflits de modélisation de la troisième dimension*

La troisième dimension peut être modélisée de trois manières [de Cambray 94] :

- le **2.5 D** associe à chaque couple (x, y) un unique z (z = f(x, y)), tel le modèle numérique de terrain, …),
- le **2.75D** étend le 2.5D pour prendre en compte l'épaisseur selon l'axe : un intervalle d'altitude [z_min, z_max] est associé à (x,y),
- le **3D** considère les points et les formes en trois dimensions (cubes, parallélépipèdes, cônes, …), ce qui permet de représenter à la fois le sol (prairie, route,…), le sur-sol (habitation, arbre,…) et le sous-sol (tunnel,…).

La figure 23 illustre ces différences de modélisations pour un pont.

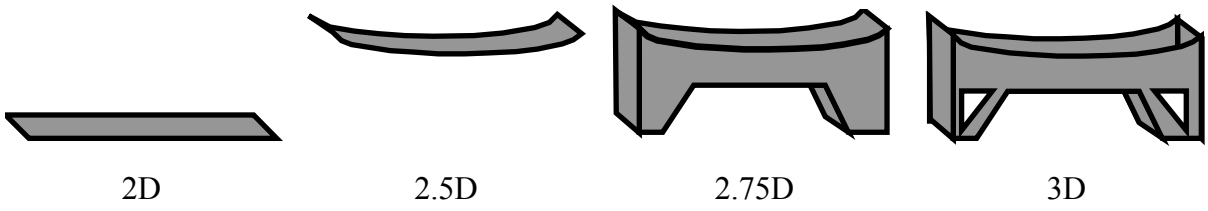



figure 23 : Modélisation de la troisième dimension, exemple d'un pont

Il y a donc **conflit de modélisation de la troisième dimension** quand la modélisation de celle-ci est différente entre les bases à intégrer. Une première solution a été proposée par [de Cambray 94] [de Cambray et Yeh 94].

##### *3.2.2.3.2 Conflits d'abstraction de la troisième dimension*

Les bases de données utilisant la même modélisation de la troisième dimension, peuvent aussi être en conflits quand les abstractions d'un même phénomène du monde réel sont différentes.

Par exemple, il existe un conflit d'abstraction de la troisième dimension pour la représentation des habitations entre la BD TOPO et le Cadastre (figure 24). En effet, la surface en 2D représentant les habitations est définie pour des hauteurs différentes. Pour la BD TOPO, comme les sources de saisie sont des photos aériennes, la surface représentant l'habitation est l'emprise de l'habitation à la hauteur des gouttières. Par contre, pour le Cadastre, la surface représentant l'habitation est l'emprise au sol (les sources de saisie étant des relevés de terrain).

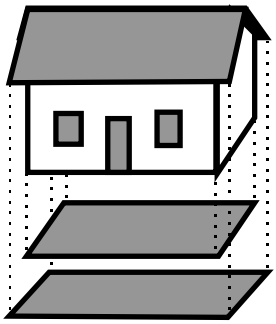



figure 24 : Abstraction de la troisième dimension pour une habitation

Ces conflits d'abstraction de la troisième dimension vont provoquer des conflits de données (3.2.6). Les BDG doivent donc contenir des spécifications de saisie (méta-données) décrivant cette abstraction de la troisième dimension afin que l'utilisateur puisse en tenir compte lors de la confrontation des données.

#### 3.2.2.4 Conflits de mode de représentation de la géométrie

La géométrie peut être stockée selon deux modes (figure 25) :

- Les structures en mode **matriciel** encore appelée **raster** ou **maillé** sont fondées sur un quadrillage régulier du terrain. Chaque nœud ou pixel du quadrillage est identifié par le numéro de sa ligne et de sa colonne. Des partitions irrégulières sont aussi proposées.
- Les structures en mode **vecteur** par contre, sont basées sur des **primitives géométriques**. Elles sont principalement le point, le segment (deux points reliés), la ligne (ensemble de segments mis bout à bout) et la surface (ligne fermée).

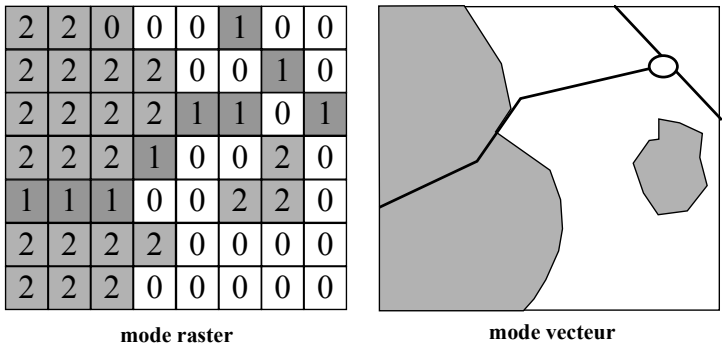



figure 25 : Mode de représentation

Ces deux modes n'ont pas les mêmes avantages. Le mode vecteur permet de représenter des objets, bien définis, partageant la même description. Par contre, le mode raster est plus adapté pour représenter des champs continus (altitudes, températures, …).

Il y a donc un **conflit de mode de représentation,** quand le mode de représentation des bases à intégrer est différent.

Pour résoudre ce type de conflits, plusieurs algorithmes de conversion ont été proposés ([Peuquet 81 a] [Peuquet 81 b]). [Piwowar et al. 90] décrit les principaux algorithmes proposés de conversion vecteur-raster et réciproquement. Hélas, ces algorithmes dégradent la géométrie des objets. Une autre solution consiste à conserver, ces deux modes de stockage [Günter 89] pour le même objet, afin de ne pas dégrader la géométrie. Une troisième piste

consiste à définir un modèle englobant le mode raster et le mode vecteur (modèle canonique) [Ramirez 97] [Egenhofer et al. 89].

#### 3.2.2.5 Conflits de méta-données géométriques

La géométrie des données des BDG dépend aussi de méta-données spécifiques (précision, résolution, exactitude).

La **précision** désigne l'unité de mesure de la géométrie, la précision d'une BDG peut être le mètre, le centimètre. Il y a un **conflit de précision** si l'unité de mesure de la géométrie des deux bases est différente.

Les **résolutions** géométriques désigne généralement la taille du plus petit objet représentable. Elle peut également être associée à un ensemble de critères.
Pour les objets surfaciques, les principaux critères sont (figure 26) :

- la surface minimale (1),
- la section minimale (2),
- la longueur d'arête minimale (3),
- l'inter-distance minimale (4).

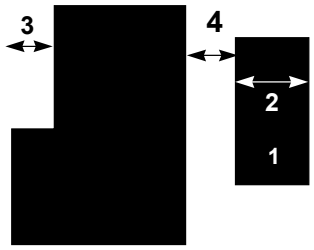



figure 26 : Critères définissant la résolution
pour des objets surfaciques.

Pour les objets linéaires, la résolution peut être définie par :

- la longueur minimale,
- la longueur d'arête minimale,
- l'inter-distance minimale.

Il existe donc un **conflit de résolution** entre deux bases à intégrer, quand un des critères définissant la résolution est différent [Shepherd 92]. Ces critères peuvent être définis classe par classe ou couple de classes par couple de classes. Par exemple, l'inter-distance minimale entre deux objets de la classe HABITATION peut être de 1 m alors que l'inter-distance minimale entre un objet de la classe HABITATION et un objet de la classe TRONÇON_ROUTE est de 0 m.

L**'exactitude** désigne l'écart entre la mesure stockée dans la BDG et une mesure parfaite, qui ne serait entachée d'aucune erreur. Du fait des processus de saisie différents, des **conflits d'exactitude** de la géométrie apparaissent [Laurini 96].

Ces trois types de conflits vont provoquer des conflits de données (3.2.6). Des phénomènes du monde réel représentés dans la BDG 1, ne seront pas représentés dans la BDG 2 du fait de la résolution. De même, la géométrie de données représentant le même phénomène du monde réel, pourra différer. Ce conflit sera abordé au niveau de l'intégration des géométries (4.3.6.2).

#### 3.2.2.6 Conflits de modélisation de la topologie

Les BDG ne se distinguent pas seulement par la localisation de leurs instances, mais aussi par les contraintes topologiques et les relations topologiques qui décrivent les liens entre les géométries de leurs instances (partage de géométrie, adjacence, frontière, extrémité, …). Ces relations topologiques sont des relations prédéfinies, obligatoires formant un modèle topologique. Plusieurs modèles ont été définis. Ils peuvent être groupés en trois familles (annexe 7.1.3.3) :

- les modèles topologiques de **graphe** ou de **réseau,**
- les modèles topologiques de **carte** ou de **surface,**
- les modèles **spaghetti** (sans topologie).

Il y a un **conflit de modélisation de la topologie** quand les modèles topologiques des BDG à intégrer sont différents. Trois **solutions** sont possibles :

- l'ajout des relations topologiques dans la base la moins riche [Ubeda et Egenhofer 97]. Malheureusement, la construction de la topologie a posteriori, va modifier les géométries (un noeud doit être créé lors de l'intersection de deux arcs pour un modèle de graphe planaire), ce qui génère des imprécisions qui ne peuvent être résolues automatiquement,
- le développement de modèles dans lesquels des données ayant des topologies différentes peuvent être stockées (les données ayant des topologies différentes ne sont pas fusionnées) comme $GéO_2$ [David et al. 93 c]),
- la définition d'un modèle permettant d'exprimer la topologie selon différentes résolutions [Puppo et Dettori 95] [Bertolotto et al. 94].

### *3.2.3 Conflits de définition des classes*

La classification selon un ensemble de critères est indispensable pour représenter et manipuler les phénomènes du monde réel aisément [Booch 91]. Cette opération est particulièrement difficile pour les BDG. En effet les phénomènes du monde réel sont pour la plupart des **phénomènes continus** et **hétérogènes** (valeurs différentes sur l'ensemble de leur emprise pour leurs attributs) avec des contours qui sont mal définis. Or, les instances des classes sont des objets de valeurs **homogènes** et **discrètes** avec des **limites précises**.

Les conflits liés à la définition des classes sont donc nombreux. Nous les avons regroupés en trois groupes : les conflits de classification (3.2.3.1), les conflits de spécification (3.2.3.2), les conflits de fragmentation (3.2.3.3)).

#### 3.2.3.1 Conflits de classification

Les conflits de classification apparaissent, lorsque deux classes sémantiquement liées peuvent décrire des phénomènes du monde réel différents, c'est-à-dire lorsqu'un même ensemble de phénomènes du monde réel est classé différemment d'un schéma à l'autre. Ces conflits sont présents dans les BD classiques : ils ont été décrits dans [Kim et al. 93] [Parent et Spaccapietra 96]. Par exemple, les routes (la N7, la route Napoléon) et les itinéraires de grande randonnée (GR 20), sont classés différemment dans la BD CARTO et dans la BD TOPO :

- la BD CARTO utilise comme critère de classification l'appartenance ou non à un classement administratif. Les routes ayant un numéro administratif (la N7) sont regroupées dans la classe ROUTE, les routes nommées (la route Napoléon) et les GR (le GR 20) sont regroupés dans la classe ITINERAIRE_ROUTIER,

- la BD TOPO utilise comme critère de classification le mode de déplacement, une classe ROUTE est définie pour les routes (la N7, la route Napoléon) et une classe GR pour les GR (GR 20).

Trois sous-classes de conflits de classifications complexes vont maintenant être décrites (conflits de regroupements, conflits de résolution, conflit de données / méta-données).

#### *3.2.3.1.1 Conflits de regroupements*

Pour des classifications utilisant le même critère de classification, un conflit de classification peut exister si l'interprétation de ce critère donne lieu à **des regroupements différents.**

Par exemple, pour l'intégration d'une base de données sur la gestion de l'eau (base eau) et d'une base de données sur la gestion de la forêt (base forêt) [Gouvernement du Québec 92] ayant chacune les classes « zone humide » et « zone boisée», un conflit de regroupements apparaît (figure 27). En effet, le critère de classification (le type d'occupation du sol) dans les deux bases est identique, mais l'interprétation de ce critère en zones humides boisées diffère selon les bases : une zone humide boisée est représentée par une instance de la classe ZONE HUMIDE pour la base eau, et par une instance de la classe ZONE BOISEE pour la base forêt. Les zones humides boisées ne font donc pas partie de la même classe.

<table>
<tr><th></th><th>Zone humide</th><th>Zone humide boisée</th><th>Zone boisée</th></tr>
<tr><td>Base eau</td><td colspan="2">zone humide</td><td>zone boisée</td></tr>
<tr><td>Base forêt</td><td>zone humide</td><td colspan="2">zone boisée</td></tr>
</table>

figure 27 : Exemple de conflit de classification dû au contexte

#### *3.2.3.1.2 Conflits de résolution*

Pour chaque critère de classification, il faut fixer son **niveau de détail** ou **d'abstraction**. Deux concepteurs de bases de données peuvent donc choisir les mêmes critères pour définir leur classification mais avec un niveau de détail différent. Nous sommes en présence d'un conflit de résolution. Par exemple, pour deux BDG décrivant l'occupation du sol [Rigaux et Scholl 95], le critère de classification sémantique choisi dans la première base, distingue les classes ZONE_BATIE et ZONE_CULTIVEE tandis que le critère de classification sémantique choisi dans la seconde base, est plus détaillé et distingue les cinq classes : CEREALE, FLEUR, VIGNE, ZONE_URBAINE et ZONE_PAVILLONNAIRE.

Ce type de conflit peut être rencontré pour un critère de **résolution géométrique**. Par exemple, dans la BD CARTO, il existe une seule classe pour les tronçons de rivière : TRONCON_HYDROGRAPHIQUE, en revanche dans la BD TOPO, il existe deux classes TRONCON_COURS_D'EAU pour les tronçons de largeur inférieure à 7,5 mètres et SURFACE_COURS_D'EAU pour les autres. Ce type de conflits est spécifique des BDG.

#### *3.2.3.1.3 Conflit de données / méta-données*

La classification des objets est réalisée dans un premier temps lors de la définition des classes puis lors de la définition des attributs [Kent 81]. Pour une instance, le nom de sa classe est une méta-donnée alors que la valeur d'un de ses attributs est une donnée.

**Un conflit de données / méta-données** survient donc lorsqu'une information de classification correspond à une donnée dans une base et à une méta-donnée dans l'autre base [Saltor et al. 92]. Par exemple, dans la BD CARTO, si une instance de la classe TRONÇON DE

ROUTE prend la valeur « sentier » pour l'attribut *état physique de la route*, ce tronçon est un sentier. Dans la BD TOPO, ce même objet serait une instance de la classe SENTIER. La collection des sentiers forme donc une classe dans la BD TOPO, alors qu'elle est le résultat d'une sélection dans la BD CARTO.

#### *3.2.3.1.4 Solutions pour les conflits de classification*

Plus de dix techniques sont envisageables pour l'intégration de deux classes [Dupont 95 b] (Annexe 7.3.1). Le choix de l'une d'entre elles est bien entendu fonction de l'objectif retenu pour la BD intégrée et des relations ensemblistes qui relient leurs instances. Nous ne prendrons pas en compte les techniques de multi-instanciation, car cette technique est implémentée uniquement dans quelques prototypes (COCOON [Scholl et Schek 90] et Dual [Perl et al. 89] par exemple).
Pour résoudre les conflits de classification, les techniques suivantes peuvent être utilisées :

- la **fusion** (7.3.1.2) qui consiste à créer dans le schéma intégré, une classe ayant pour attribut, l'ensemble des attributs des classes à intégrer et pour instances, l'union des instances.
- la **partition** (7.3.1.5) qui consiste à créer une classe pour chaque intersection et pour chaque différence. Les attributs des intersections sont l'union des attributs. Les attributs des différences sont les attributs de la classe initiale.
- la **généralisation - spécialisation** (7.3.1.10) qui définissent des relations d'héritage entre les classes à intégrer [Larson et al. 89] [Gotthard et al. 92].

Par exemple, pour intégrer la base eau et la base forêt (figure 27) :

- la fusion donnera une unique classe ZONE_HUMIDE_OU_BOISEE,
- la partition donnera 3 classes ZONE_HUMIDE_NON_BOISEE, ZONE_HUMIDE_BOISEE et ZONE_BOISEE_NON_HUMIDE.
- la généralisation - spécialisation donnera les classes suivantes :
    - ZONE_HUMIDE_OU_BOISEE,
    - ZONE_HUMIDE qui hérite de ZONE_HUMIDE_OU_BOISEE,
    - ZONE_BOISEE qui hérite de ZONE_HUMIDE_OU_BOISEE,
    - ZONE_HUMIDE_ET_BOISEE qui hérite à la fois de ZONE_HUMIDE et de ZONE_BOISEE.

Ces techniques peuvent aussi être combinées entre elles (voir les autres opérations de l'annexe 7.3.1).

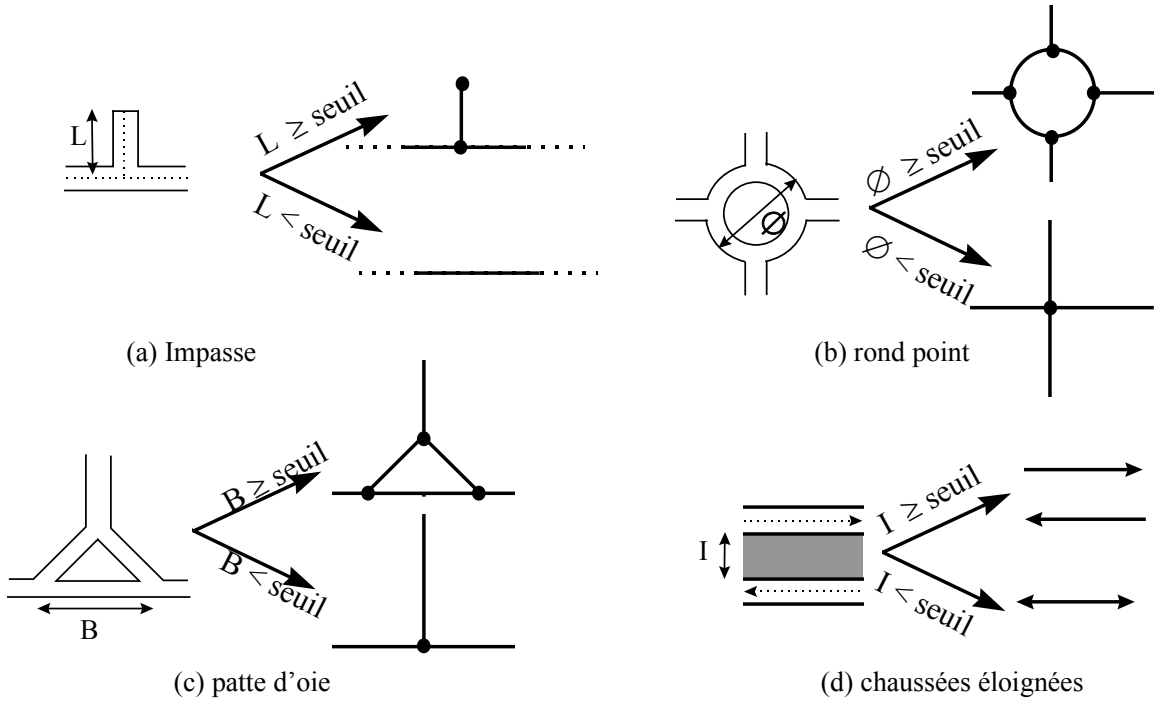
### 3.2.3.2 Conflits de critères de spécification

Les phénomènes du monde réel étant des phénomènes continus et de tailles variables, la représentation de ceux-ci par une instance d'une classe ou d'une autre peut poser des problèmes. De même, il n'est pas évident de savoir si un phénomène fait partie de l'univers de la BDG. Pour répondre à ces problèmes, des **spécifications de contenu et de saisie** ont été définies pour les BDG. Elles dépendent des méta-données sur la géométrie (précision, résolution, exactitude).

#### *3.2.3.2.1 Conflits de critères de sélection*

Les **critères de sélection** déterminent si un phénomène doit être représenté par un objet de cette classe. Par exemple, la BD TOPO a sélectionné comme instances de la classe TRONÇON_ROUTE toutes les rues ouvertes au public de plus de 50 mètres (tableau 3) et éliminé toutes les impasses de moins de 50 mètres.

Il existe un **conflit de critères de sélection** quand pour deux classe en correspondance un des critère de sélection prend une valeur différente (figure 28 a).




(a) Impasse (b) rond point

(c) patte d'oie (d) chaussées éloignées

figure 28 : Représentations d'éléments routiers en fonction des seuils

| **Spécification** | BD TOPO | GEOROUTE | BD CARTO V2 |
|---|---|---|---|
| tronçon quelconque | tous tronçons publics (route, chemin, allée, sentier) | tronçons carrossables + chemins ayant un toponyme et des habitations | **zone agglomérée** : tronçons du réseau principal **autre** tous |
| impasse | L>50 m L>100 m si maison isolée | L> 10 m | L>1000 m L>200 m si maison L > 500 m si littoral |
| rond-point | ∅ > 25 m | ∅ > 30 m | ∅ > 100 m |
| pattes d'oie | écartement > 50 m | écartement > 20 m | non saisie |
| chaussées éloignées | intervalle > 25 m ou Dénivelé > 1 m | intervalle > 20 m | intervalle > 100 m |

tableau 3 : Tableau de quelques critères de spécification des BD de l'IGN

#### *3.2.3.2.2 Conflits de critères de décomposition*

Les **critères de décomposition** permettent de définir à partir de quel **seuil** un phénomène du monde réel sera représenté par un ou plusieurs objets. Par exemple, dans la BD TOPO, un carrefour est considéré comme un rond-point (ensemble de noeuds routiers reliés par des tronçons) si son diamètre est supérieur à 25 mètres et comme un carrefour simple (un noeud routier) dans le cas contraire (figure 28 b). Des critères similaires existent pour définir à partir de quel écartement, la base individualise chacune des pattes d'oie (figure 28 c) ou à partir de

quel intervalle entre les chaussées, les chaussées sont dites « éloignées » et donc représentées par plusieurs tronçons parallèles (figure 28 d).

Il existe donc **un conflit de critère de décomposition** entre deux classes en correspondance quand un des critères de décomposition a une valeur différente.

Le tableau 3 montre qu'il existe un grand nombre de conflits de critère de spécification entre les BDG de l'IGN pour le thème routier. Ces conflits vont entraîner des conflits au niveau des données (données sans correspondants,…) (3.2.6), des conflits de classification (3.2.3.1) et des conflits de fragmentation (3.2.3.3). Ils seront donc traités lors de la résolution des conflits « avals » qu'ils entraînent.

### 3.2.3.3 Conflits de fragmentation

Les abstractions des phénomènes en objets peuvent fragmenter un phénomène en plusieurs objets suivant les valeurs des attributs. Le découpage en objets n'est donc pas identique d'une base à l'autre. Il existe donc un **conflit de fragmentation** [Dupont 95a] quand un objet dans une base correspond à plusieurs objets dans l'autre base ou, quand un ensemble d'objets dans une base correspond à un ensemble d'objets dans l'autre base sans qu'il soit possible d'établir une relation bijective entre les objets. Il est à noter que les ensembles d'objets peuvent être de la même classe ou de classes différentes.

#### *3.2.3.3.1 Conflits de segmentation*

Le conflit de segmentation est un conflit de **découpage en objets homogènes**, **selon des attributs différents**. Par exemple, les routes peuvent être segmentées en tronçon de routes selon le nombre de voies dans la première base et segmentées en tronçon de routes selon le revêtement dans la seconde base (figure 29).

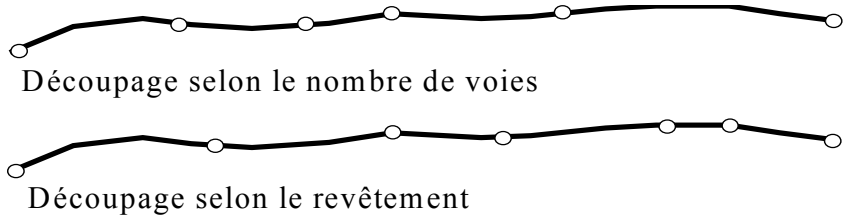



figure 29 : Exemple de conflit de segmentation

#### *3.2.3.3.2 Conflits de granularité*

Des conflits de fragmentation peuvent aussi survenir quand les attributs sont identiques, car pour chaque attribut, une limite inférieure pour la taille des objets est définie. Un objet homogène est créé uniquement si la valeur des attributs est constante sur une longueur ou une surface minimale, Cette longueur ou surface est appelée **granularité de l'attribut**.

Il existe donc un **conflit de granularité** entre deux classes, si pour un attribut en correspondance la granularité est différente. Par exemple, pour les classes TRONÇON_ROUTE, la granularité est différente entre GÉOROUTE et la BD CARTO. Pour GÉOROUTE, la valeur des attributs doit être constante sur 10 m au minimum ; par contre pour la BD CARTO, la valeur des attributs doit être constante sur 1 000 m au minimum. Ainsi, trois tronçons de GÉOROUTE peuvent correspondre à 1 tronçon de la BD CARTO, si le deuxième tronçon GÉOROUTE est trop petit pour la granularité de la BD CARTO (figure 30).

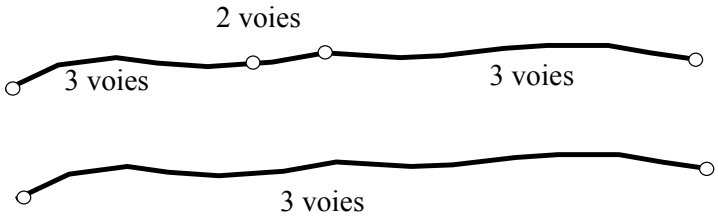



figure 30 : Exemple de conflit de granularité

Brugger introduit une notion similaire à la granularité : l'**homogénéité**[9] [Brugger 94] [Brugger 95]. Les conflits qui découlent de critères d'homogénéité différents sont similaires aux conflits de granularité. Cependant, cette contrainte d'homogénéité n'a pas été rencontrée dans les BDG.

#### *3.2.3.3.3 Conflits de décomposition*

Un conflit de décomposition intervient quand un objet vu comme un tout dans une base est décomposé en plusieurs objets dans l'autre base. Ces objets peuvent être des instances de la même classe ou de classes différentes. Par exemple, dans la figure 31, les routes de la BD1, sont décomposées en tronçons de route alors que pour la BD 2, les routes sont modélisées en tronçons de chaussée (ensemble de voies allant dans la même direction) et en séparateurs


figure 31: Exemple de conflit de décomposition pour une route

#### *3.2.3.3.4 Solutions pour les conflits de fragmentation*

Quelques solutions ont été proposées pour répondre à ce type de conflits.

Pour les conflits de segmentation, la **segmentation dynamique** proposée par [Maguire et al. 92] et applicables aux objets linéaires orientés, consiste à définir la valeur d'un attribut en fonction de la distance au point initial. Cette solution a été implementée sous Arc/Info®. Une solution équivalente a été proposée pour les objets surfaciques, mais elle n'a pas été implementée car il n'y a pas de relation d'ordre sur une surface.

Pour les conflits de granularité, aucune solution n'a été proposée.

Pour les conflits de décomposition, la solution standard consiste à créer des **relations de composition** entre les objets en conflits [Dupont 95b]. Par exemple, un tronçon de route est composé de tronçons de chaussées et de séparateurs.

### *3.2.4 Conflits de structures*

Nous avons regroupé dans les conflits de structure, les conflits de structures classiques et les conflits de stockage de l'information.

---

[9] l'homogénéité est définie comme le pourcentage minimum de l'emprise du phénomène du monde réel vérifiant les valeurs des attributs et l'appartenance à la classe de l'objet homogène associé.

### 3.2.4.1 Conflits de structures classiques

Un conflit structurel survient lorsque les éléments en correspondance sont décrits par des concepts différents [Kim et al. 93] [Shoval et Zohn 91]. Trois concepts seront ici distingués : la **classe**, **l'attribut** et la **relation**.

Par exemple, les tronçons de bacs sont représentés par des instances de la classe TRONÇON DE BAC dans la BD CARTO alors qu'ils sont matérialisés par l'attribut *présence d'équipements spéciaux* dans GEOROUTE (figure 32).

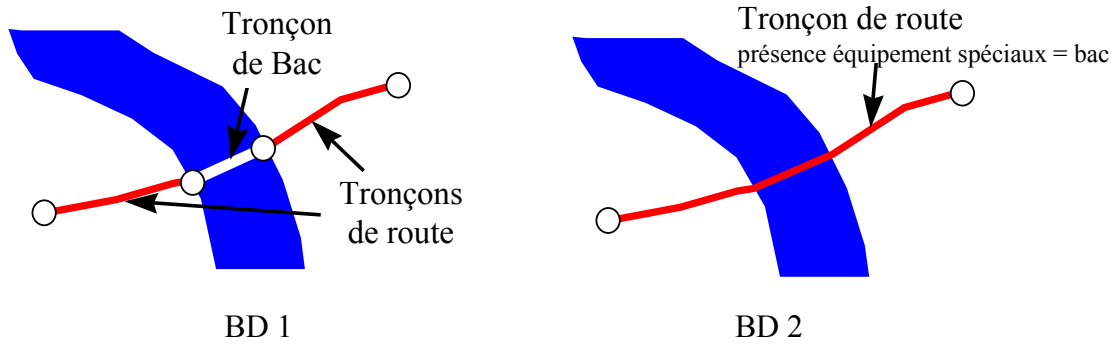



figure 32: Exemple de conflit de structure entre une classe et un attribut

De même, les ponts sont représentés par des objets de la classe PONT dans la BD 1 alors qu'ils sont représentés par des relations entre des objets de classes ROUTE et RIVIERE dans la BD 2 (figure 33) (conflit de structure classe / relation).

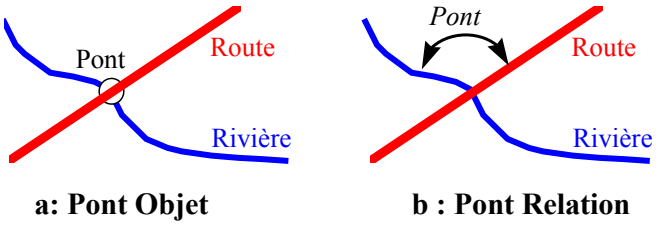



figure 33 : Exemple de conflit de structure classe / relation

Les **solutions** proposées pour les conflits de structure, consistent à choisir parmi les structures en conflit, la structure la moins contrainte [Spaccapietra et Parent 91], c'est-à-dire la structure ayant la capacité de décrire les instances des éléments initiaux. Ainsi, les populations des éléments en correspondance pourront être représentées dans le schéma intégré.

Pour les conflits de structure **classe / attribut** (figure 34 a), la structure de classe sera choisie, dans le schéma intégré. L'attribut sera remplacé par une classe et une relation entre cette nouvelle classe et la classe de l'ancien attribut sera créée.

Pour les conflits **classe / relation** (figure 34 b), la structure de classe sera choisie, dans le schéma intégré. La relation sera remplacée par une classe et deux relations entre cette nouvelle classe et les classes reliées à l'ancienne relation.

Pour les conflits de structure **relation / attribut** (figure 34 c) la structure de relation sera choisie, dans le schéma intégré. Une relation représentant la relation et l'attribut en correspondance sera créée. Elle reliera les classes initialement reliées à la relation en correspondance et à la classe de l'attribut en correspondance.

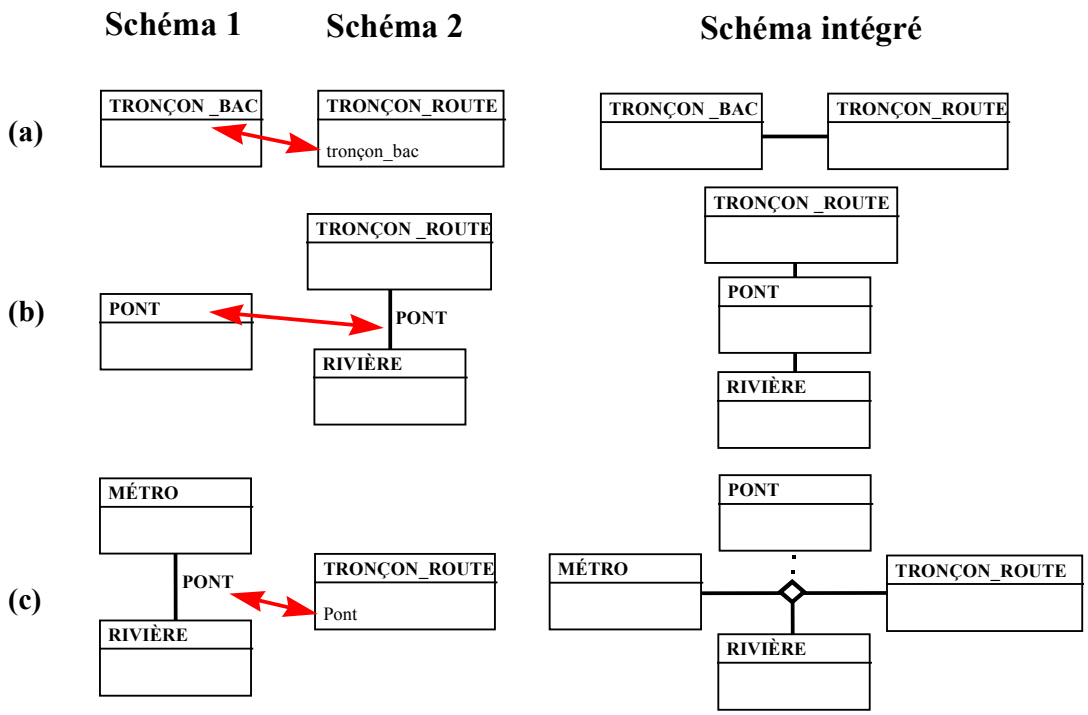



figure 34 : Solutions pour les conflits de structure

#### 3.2.4.2 Conflits de stockage de l'information

Aux conflits de structures classiques, s'ajoutent les conflits de stockage de l'information. Une des particularités des BDG est la quantité d'informations **implicites** [Grumbach et al. 96] que l'on peut déduire (relation de proximité, relation d'inclusion, objets, propriétés de l'objet). Les concepteurs de la BDG peuvent donc choisir de matérialiser des informations dans la base ou de leur laisser un caractère implicite.

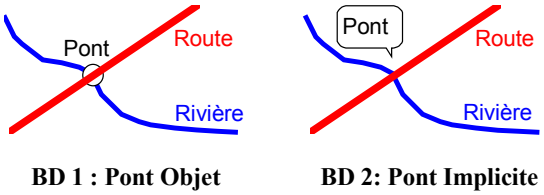



figure 35 : Exemple de conflit de stockage

Il y a **conflit de stockage** quand une information stockée dans une base de données géographiques correspond à une information qui doit être déduite dans l'autre base. Par exemple, dans la figure 35, dans la BD 1, les ponts sont matérialisés comme des objets alors que dans la BD 2, les ponts doivent être déduits du croisement d'une rivière et d'une route.

### *3.2.5 Conflits de description sémantique et géométrique*

Les conflits de description sémantique et géométrique résultent de différences entre les propriétés (attributs, méthodes) des classes en correspondance [Larson et al. 89] [Kim et al. 93].
Ces différences peuvent porter sur le type des classes, leurs attributs descriptifs. Les conflits de description sur les attributs géométriques seront aussi abordés.

#### 3.2.5.1 Conflits de description des classes

Un conflit de description peut être lié à la classe dans sa **totalité**. Effectivement, les classes sont décrites par leur nom, leurs identifiants et par l'ensemble de leurs attributs. Un conflit portant sur la description des classes survient entre deux classes, quand au moins un de ces critères est différent. Par exemple, les classes EMPLOYÉ de la BD 1 et OUVRIER de la BD 2 sont en conflit de description (nom différent).

#### 3.2.5.2 Conflits de description simples entre les attributs

Les conflits de description peuvent se situer au niveau des attributs. Ce problème a particulièrement été détaillé dans [Larson et al. 89]. Ils caractérisent chacun des attributs en fonction de sa structure, le domaine de valeur, l'échelle (l'unité de mesure employé pour le domaine), …

En général, les taxonomies portant sur les différences de description ne traitent que des conflits décrits ci-dessus entre deux attributs en correspondance.

Pour ces conflits, des solutions ont été proposées uniquement pour les problèmes « simples » (renommage pour les conflits de nom, …).

#### 3.2.5.3 Conflits de description n-aires entre attributs

Les conflits de description n-aires sont plus complexes. Ils surviennent entre les attributs lorsque l'information contenue par un attribut correspond à l'information apportée par plusieurs attributs ou lorsque l'information contenue par plusieurs attributs correspond à l'information apportée par plusieurs attributs, sans qu'il soit possible de déterminer des équivalences plus fines entre les attributs.

Ces conflits n-aires sont souvent dus à des regroupements d'informations proches ayant une partie de l'information qui peut être déduite de l'autre, ou à des attributs incompatibles. Dès lors, l'information obtenue à l'aide des ensembles d'attributs est en correspondance, mais les domaines de valeurs pris un à un ne sont pas compatibles.

| |
|---|
| **Tronçon de route BD CARTO V2**<br>**Etat physique** : enum (revêtue, non revêtue, en construction, chemin exploitation, sentier)<br>**Vocation liaison** : enum (autoroute, grande circulation, liaison locale, bretelle, piste cyclable) |
| **Tronçon de route Géoroute**<br>**Etat physique** : enum (revêtue, non revêtue, en construction)<br>**Vocation liaison** : enum (autoroute, artérielle, distribution; desserte, bretelle, chemin ou sentier, passerelle, escalier, voie rapide urbaine) |

tableau 4: Exemple de conflits de description n-aires portant sur les domaines des attributs.

Par exemple, il existe un conflit de description n-aire entre les deux attributs *Etat physique* et *Vocation liaison* de la classe TRONÇON_ ROUTE de la BD CARTO V2 et les deux attributs *Etat physique et Vocation liaison* de la classe TRONÇON_ROUTE de GEOROUTE (tableau 4). Ainsi, un chemin non revêtu aura pour valeurs :

- dans la BD CARTO V2 :
    - Etat physique = « chemin d'exploitation »
    - Vocation liaison = « liaison locale »
- dans GEOROUTE
    - Etat physique = « non revêtu »
    - Vocation liaison = « chemin ou sentier »

Ce genre de conflits n'a pas été traité jusqu'à présent, il rend particulièrement ardu l'intégration des attributs des BDG.

#### 3.2.5.4 Conflits de description géométrique pour les données vecteurs

Un objet géographique est relié à une géométrie pour décrire sa localisation et sa forme. Pour les données vecteurs, la géométrie des phénomènes du monde réel est représentée par des primitives géométriques (point : dimension 0, ligne : dimension 1, surface : dimension 2).

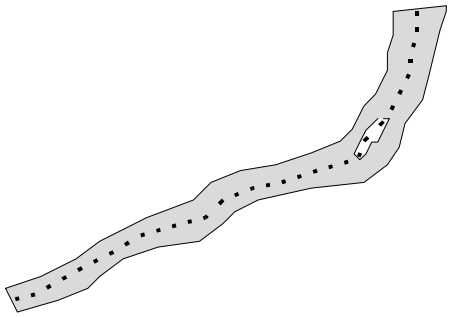


figure 36 : Exemple de conflit de dimension de la géométrie

Un **conflit de description géométrique** survient quand des primitives géométriques de dimensions différentes représentent un même phénomène du monde réel [Laurini 96]. Par exemple, pour un fleuve frontière (figure 36), sa géométrie sera surfacique dans la base hydrographique et linéaire dans la base administrative. Ce conflit résulte du **type d'application.**

Le choix de la dimension de la géométrie peut aussi être dû à la **différence d'échelle**[10] entre les bases à intégrer. Par exemple (figure 37), les rivières surfaciques dans une base à grande échelle peuvent correspondre à des rivières linéaires dans une base à une échelle plus petite. De même, les habitations surfaciques peuvent correspondre à des habitations ponctuelles et des barrières linéaires peuvent correspondre à des barrières ponctuelles.

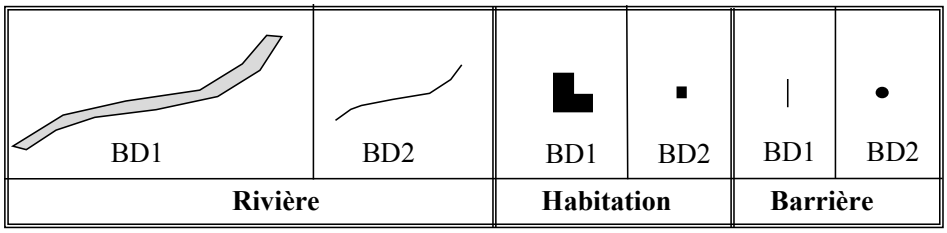



figure 37 : Exemple de conflit de dimension de la géométrie

Ce type de conflit est propre aux BDG. Ils sont dus principalement à la différence de **granularité** des attributs géométriques. Cependant, pour les attributs temporels, des conflits

[10] précision, résolution, exactitude

semblables existent [Montanari et al. 92] [Euzenat 94] [Euzenat 95]. Un attribut temporel de type intervalle (dimension 1) peut correspondre à un attribut temporel de type instant (dimension 0). Pour les BDG, les **solutions** proposées consistent à définir des méta-données précises, permettant de connaître pour chaque donnée la dimension de sa géométrie [Stephan et al. 93], ou à définir une structure permettant de relier les géométries des données à différentes échelles [Puppo et Dettori 95] [Timpf et Frank 95].

### *3.2.6 Conflits de données*

Ce dernier type de conflit survient lorsque les objets en correspondance ont **des valeurs différentes pour les attributs en correspondance**.

Dans les bases de données classiques, les causes d'un conflit de données peuvent être les erreurs de saisie, des sources d'information différentes, des versions différentes, des mises à jour différées, …

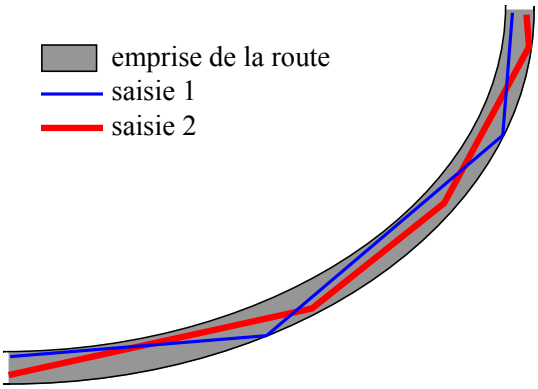



figure 38 : Exemple de saisies différentes pour une même route

Pour les BDG, les conflits de précision, de résolution, d'exactitude et de spécification vont aussi entraîner des conflits de données. Les processus de saisie vont aussi provoquer des conflits de données. En effet, pour les objets des bases de données classiques, les valeurs des attributs sont « faciles » à déduire. Par contre, pour les objets géographiques, la valeur de l'attribut géométrique est plus difficile à évaluer avec précision et plusieurs valeurs sont possibles. Deux saisies manuelles de la géométrie produiront deux restitutions différentes. Par exemple, pour la figure 38, les deux saisies de la route effectué à partir d'une photographie aérienne, sont différentes.

Ces conflits portant sur la géométrie peuvent aussi être engendrés par le **caractère flou** des limites des objets géographiques [Shepherd 92], par exemple, la limite d'une forêt est mal définie.

De même, des **conflits comme les conflits de résolution** provoquent des conflits de données. Par exemple, un tronçon linéaire composé de petits segments de longueur inférieure à la longueur minimale pour cette résolution sera représenté par une ligne composée de segments plus grands.

Des opérations de **généralisation** peuvent entraîner des conflits de données plus complexes. Par exemple, la structuration (figure 39) permet de résoudre les problèmes de résolution en diminuant le nombre d'instance d'une même classe de même valeur dans une zone. Cette opération de simplification spatiale qui consiste à conserver l'expressivité globale **(**la structure matricielle pour les habitations de la figure 39) au dépend de la cardinalité (le nombre d'habitations) et de l'expressivité élémentaire. Un des objets structurant en lui-même n'a

aucune signification : l'ensemble, par sa forme, l'espacement entre ses objets et la position de ses objets, permet de traduire l'aspect général des objets de départ. Il n'existe pas de correspondance entre un objet structurant et un objet structuré. Par contre, il existe une correspondance entre un ensemble d'objets A, et un ensemble d'objets B. Cette opération crée donc un conflit **n-aire** au niveau des données.

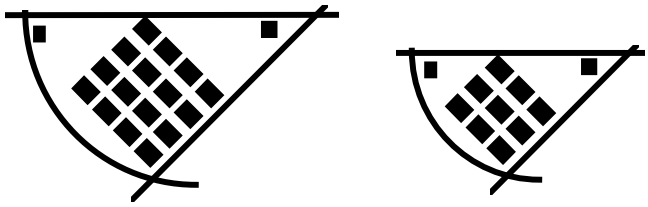



figure 39 : Exemple de structuration d'habitations

Du fait des nombreux conflits de données rencontrés dans les BDG, les valeurs des objets en correspondance vont souvent différer et a fortiori, l'identification des objets sémantiquement équivalents va être rendue plus délicate que pour les données des BD classiques. Afin de résoudre cette difficulté, des méthodes d'appariement évoluées ont été développées.

### *3.2.7 Conclusion sur la taxonomie des conflits d'intégration de BDG*

Cette taxonomie des conflits d'intégration recense 6 classes de conflits d'intégration de BDG (tableau 5). Certains de ces conflits sont propres aux BDG, les méthodes d'intégration de BD classiques doivent donc être étendues et complétées. Elles devront englober les solutions présentées dans la taxonomie et proposer de nouvelles techniques pour les conflits non encore résolus.

- **conflits de sources de données**
- **conflits d'hétérogénéité**
    - conflits de modèle
    - conflits de type de positionnement
    - conflits de gestion de la troisième dimension : la hauteur
        - conflits de modélisation de la troisième dimension
        - conflits d'abstraction de la troisième dimension
    - conflits de mode de représentation de la géométrie
    - conflits de méta-données géométriques
    - conflits de modélisation de la topologie
- **conflits de définition des classes**
    - conflits de classification
        - conflits de regroupements
        - conflits de résolution
        - conflit de données / méta-données
    - conflits de critère de spécification
        - conflits de critères de sélection
        - conflits de critères de décomposition
    - conflits de fragmentation
        - conflits de segmentation
        - conflits de granularité
        - conflits de décomposition
- **conflits de structures**
    - conflits de structures classiques
    - conflits de stockage de l'information
- **conflits de description sémantique et géométrique**
    - conflits de description des classes
    - conflits de description simples entre les attributs
    - conflits de description n-aires entre attributs
    - conflit de description de la géométrie pour les données vecteurs
- **conflits de données**

tableau 5 : Conflits d'intégration de BDG

## 3.3 Conclusion sur l'approche formelle

Cette approche formelle a permis de choisir une **méthode d'intégration** classique. Les conflits **d'intégration** des BDG, ont aussi été présentés. Pour chaque conflit, les solutions proposées ont été décrites. Ainsi, la **méthode d'intégration** classique pourra être étendue aux BDG, en intégrant des solutions déjà proposées et en développant de nouvelles solutions.

# 4. Extensions de la méthode d'intégration pour les BDG

Une fois, le processus d'intégration classique sélectionné et l'ensemble des conflits d'intégration des BDG énumérés, les enrichissements nécessaires pour le processus d'intégration de BDG peuvent être exposés.

Les enrichissements ne porteront néanmoins que sur les données **vecteur** à **deux dimensions**. L'intégration de données en mode raster et vecteur sera simplement évoqué, car ce problème a déjà été plusieurs fois abordé (3.2.2.4). L'intégration de BDG gérant la troisième dimension ou la dimension temporelle ne sera pas traité. De même, l'intégration des relations d'héritage n'a pas été attaquée, car ce problème a déjà été traité pour les BD classiques par [Mannino et al. 88] [Reddy et al. 94] et ne demande pas à être étendu pour les BDG. Cette thèse n'abordera pas non plus, l'intégration des comportements et des fonctions des SIG initiaux. Cette intégration est nécessaire pour ne pas les développer à nouveau dans le SIG intégré. Plusieurs travaux ont analysé cette tâche et proposé des solutions [Abel et al. 94 a ] [Abel et al. 94 b] [Abel et al. 92] [Kemppainen et Albrecht 96] [Voisard et Schweppe 94].

Le plan suivi dans cette partie sera le suivant. Dans le chapitre 4.1, les BDG de l'IGN à intégrer, seront présentées ainsi que le prototype de SGBD géographiques qui a servi de support ($GéO_2$). Le chapitre 4.2 décrira la première phase du processus, la pré-intégration de BDG qui permet de mettre en conformité les BDG. Dans le chapitre 4.3, pour chaque conflit spécifique au BDG non résolu lors de la pré-intégration, sera donnée une syntaxe pour sa déclaration et une technique d'intégration destinée à le résoudre en fonction de la stratégie d'intégration retenue.

## 4.1 Introduction : présentation des BDG à intégrer et de $GéO_2$

Notre processus d'intégration de BDG s'appuie sur un processus en trois phases. Il a été validé par son application expérimentale aux BDG de l'IGN, utilisant en tant que plate-forme le prototype de SIG $GéO_2$. Ces bases et ce prototype sont présentés ci-après.

### *4.1.1 Présentation des BDG et de leur intégration*

L'IGN comme la plupart des agences cartographiques, produit plusieurs cartes à différentes échelles organisées en séries et en conséquence plusieurs bases de données géographiques. Quatre bases principales sont commercialisées : la BD TOPO®, la BD CARTO®, GEOROUTE® et la BD altimétrique® (base de données du relief sous la forme de modèle numérique de terrain ou de courbes de niveau). Elles possèdent des informations correspondant à différentes fonctionnalités et différentes échelles. L'IGN a choisi de produire plusieurs bases de données car il n'existe pas dans les SIG actuels d'outils permettant de généraliser automatiquement les représentations les moins détaillées à partir de la représentation la plus détaillée ou de dériver des bases pour des applications connexes. Cependant l'intégration de ces données semble admise pour le futur. Les trois premières bases (la BD TOPO, la BD CARTO, GEOROUTE) serviront de jeux tests pour le processus d'intégration.

#### 4.1.1.1 La BD TOPO®

La BD TOPO® [Equipe BD TOPO 94] [IGN 96 b] fournit un système de référence de localisation pour les applications s'étendant du territoire d'une commune à celui d'un département. Elle se caractérise principalement par sa description détaillée de

l'environnement, par sa **précision de l'ordre d'un mètre** et sa structure topologique de réseau. Son contenu correspond globalement au contenu de la carte au 1 : 25 000. Elle permet des sorties graphiques du **1 : 5 000 au 1 : 25 000**. Elle est produite par stéréo-restitution. La source géométrique principale est une saisie photogrammétrique (photographies aériennes). Des compléments issus d'un relevé de terrain sont aussi réalisés, ils concernent uniquement les zones masquées sur les photographies aériennes et les zones où une validation sur le terrain est nécessaire.

La BD TOPO est séparée en deux couches géométriques indépendantes :
- l'altimétrie (ou alti) qui regroupe les courbes de niveau et points cotés,
- la planimétrie (ou plani) qui regroupe tous les autres objets.

La BD TOPO a été modélisée en privilégiant le concept de classe sur celui d'attribut. Elle est donc composée d'un grand nombre de classes qui ont peu d'attributs et de relations.

#### 4.1.1.2 La BD CARTO®

La BD CARTO® [Equipe BD CARTO 94] [IGN 96 a] est un plan numérique synthétique de l'ensemble du territoire pour des échelles comprises entre le **1 : 100 000 et le 1 : 500 000**. Elle est adaptée à la cartographie de synthèse et aux applications d'études de projet, de gestion d'infrastructures et d'aménagement du territoire. Elle a un rôle de référentiel au niveau départemental et régional, facilitant le partage d'informations entre les différents acteurs d'une même collectivité territoriale, et sur lequel chaque utilisateur peut rattacher les données de son domaine thématique.

Elle est constituée à partir des cartes au 1: 50 000 de l'IGN (scannage des cartes, vectorisation puis structuration topologique) et d'imagerie spatiale SPOT. La **précision** est de l'ordre de la **vingtaine de mètres**.

Les éléments du terrain sont regroupés en 9 thèmes (réseau routier, réseau ferré, hydrographie, franchissement, unité administrative, habillage, toponymes, équipement, occupation du sol). A chaque thème correspond une couche géométrique de type graphe planaire.

LA BD CARTO est en cours d'enrichissement au niveau de sa sémantique et de ses instances. Actuellement, les données saisies selon la version V1 ont été transformées au format de la version V2 et seront enrichies ultérieurement lors des mises à jour.

#### 4.1.1.3 GEOROUTE®

GEOROUTE® [Projet BD Routière 92] [GEOROUTE 94] [IGN 96 c] est une base de données d'informations géographiques routières ayant une topologie de surface. Sa vocation première est de fournir les informations nécessaires à l'organisation des déplacements en milieu urbain et interurbain. C'est donc une base de données localisées, dédiée aux applications logistiques routières ainsi qu'à la recherche d'itinéraires. Sa deuxième vocation est de permettre la réalisation de plans de communes.

Elle gère des données détaillées pour les zones urbaines (agglomérations de plus de 100 000 habitants) et reprend les données de la BD CARTO pour les autres zones. Elle prend en compte le réseau routier formé par les voies carrossables et les chemins en ville, uniquement si ceux-ci ont un toponyme et desservent des habitations. Sa **précision** est de l'ordre **de 5 à 10 m** en zone urbaine. L'ensemble des géométries de GEOROUTE est regroupé sur une couche géométrique unique. Pour le thème routier, la BD TOPO est proche de GEOROUTE. Néanmoins, GEOROUTE est plus riche au niveau des attributs sémantiques et des relations tandis que la BD TOPO est plus précise au niveau de la description de la géométrie.

#### 4.1.1.4 Les données à intégrer

Il a été choisi, dans un premier temps, d'intégrer les **données routières** sur **la zone de Marne-la-Vallée (24 14 E Lagny)** de 267 km². Le choix de ce thème est dû au rôle fondamental des données routières dans les BD de l'IGN et à leur présence dans les trois bases. Les schémas conceptuels du thème routier des trois bases sont donnés en annexe 7.6 au format UML [Booch et al. 97].

La zone de Marne-la-Vallée a été retenue, car elle regroupe un réseau routier très varié (autoroutes, ruelles de ville ancienne, rues de ville nouvelle, routes de campagne,…) et dense (985 kilomètres de réseau routier pour la BD TOPO, 389 kilomètres pour la BD CARTO et 766 kilomètres pour GÉOROUTE) sur un terrain changeant (vallée de la Marne, plateau de la Brie). Pour le thème routier de cette zone, la conception d'un prototype d'intégration robuste et générique serait une démonstration probante qu'une intégration complète des bases de l'IGN est réalisable.

### *4.1.2 Présentation de GéO$_2$*

Pour intégrer les BD de l'IGN, le prototype de SGBD géographiques **GéO$_2$** [David et al. 93c] [Raynal et al. 95] a été choisi comme plate-forme. Il repose sur le SGBD orienté objet du commerce O$_2$ [O$_2$ 91]. **L'orienté objet** est avantageux pour les BDG [David et al. 93a] [David et al. 93 b] [Abrantes **96]**, car il autorise une meilleure modélisation du monde réel par des objets complexes et la définition de types correspondant aux primitives géométriques. Qui plus est, il permet d'éviter les nombreuses jointures nécessaires pour manipuler les données avec un modèle relationnel.

GéO$_2$ intègre la plupart des fonctionnalités des SIG. Le modèle de données de GéO$_2$ [David et al. 93 c] a été défini au laboratoire COGIT. Il a été construit afin que les objets géométriques (points, lignes, surfaces) soient indépendants des objets géographiques qu'ils représentent (les tronçons de routes, les habitations,…). Chaque classe d'objets géographiques a un attribut de localisation du type *Geometry*. Ce type est défini comme le plus petit **Type Abstrait de Données** (TAD) incluant les points, les lignes et les surfaces [David 91]. Les objets d'une même classe géographique peuvent donc avoir des géométries de dimensions différentes. Réciproquement, les objets géométriques connaissent les objets géographiques qu'ils représentent.

Les géométries sous GéO$_2$ sont regroupées en **couches géométriques**. Trois types de couches géométriques sont définies : les couches « Spaghetti », les couches « réseau planaire » et les couches « carte topologique » (7.1.3.3). Ainsi, plusieurs niveaux de représentation de la topologie sont possibles pour les géométries. Une base GéO$_2$ peut contenir plusieurs domaines, chaque domaine étant une représentation. Les géométries d'un domaine sont groupées dans une ou plusieurs couches. De plus, GéO$_2$ autorise le stockage de **méta-données** dans un dictionnaire (dico). Par contre, GéO$_2$ ne gère pas la troisième dimension (altitude). Le **chargement des données** des trois bases dans GéO$_2$ a été réalisé domaine par domaine à partir de lots au format de données interne à l'IGN : FEIV (Format d'Echange Interne Vecteur) [Richard 93].

## 4.2 La pré-intégration de bases de données géographiques

La phase de pré-intégration définie pour les BDG est surchargée par rapport à la phase de pré-intégration classique pour tenir compte des spécificités des BDG. Néanmoins, elle regroupe les trois mêmes tâches : le choix du modèle commun, l'enrichissement et la normalisation.

### *4.2.1 Choix d'un modèle commun*

#### 4.2.1.1 Choix d'un modèle de données pour le schéma intégré

La première tâche de la pré-intégration est de choisir un **modèle de données** pour le schéma intégré et de déterminer le mécanisme de traduction des modèles initiaux vers le modèle commun et vice versa pour résoudre les conflits de modèle (3.2.2.1). Deux solutions sont envisageables. La première consiste à utiliser un **modèle minimal**, de façon à faciliter la traduction des données vers le modèle commun. Les utilisateurs de la BD intégrée souffrent alors de la pauvreté du modèle de données. La solution inverse consiste à utiliser un **modèle le plus riche possible**. Les utilisateurs disposent alors d'un modèle performant et l'intégration se fait sans perte d'information. Dans ce cas, les modèles orientés objet semblent être les plus adaptés. L'Open GIS Consortium [Buehler et McKee 96] et l'Ordnance Survey [Sleath et Perry 96] ont choisi cette option. Cependant, le choix d'un modèle riche rend la traduction des modèles initiaux plus ardue. Entre autres, il faut enrichir manuellement les modèles initiaux pour pouvoir les traduire correctement et pour utiliser toute la richesse des modèles orientés objet.

Dans le cadre de notre étude, le modèle orienté objet de $GéO_2$ a été choisi de facto. Cependant, un mécanisme de traduction des modèles initiaux vers le modèle commun a dû être défini. En effet, le chargement automatique des lots FEIV génère des bases ayant un schéma proche du modèle entité association. Des règles de traduction simples ont donc été rajoutées. Par exemple, les classes représentant des associations binaires sans attributs ($A \leftarrow R \rightarrow B$) ont été converties en attribut des classes reliées par ces associations (A.b et B.a).

#### 4.2.1.2 Choix d'un système de référence

Le choix d'un modèle commun n'est pas suffisant, il faut aussi choisir, un système de référence pour la BD intégrée pour résoudre les conflits de type de positionnement (3.2.2.2). Si les données ont le même système de projection, il suffit de translater les données afin de les caler sur le même point de référence. Par contre, si les systèmes de projection sont différents il faudra au préalable convertir les données selon un même système de référence. Deux solutions sont envisageables :

- un **mécanisme de traduction « à la volée »**, qui à chaque requête, traduit les données utilisées. Ce mécanisme est transparent pour l'utilisateur mais ralentit les traitements,
- une **traduction définitive** des données.

Si les données ne sont pas rattachées à un système de référence, il est nécessaire au préalable de les géo-référencer, c'est-à-dire de les rattacher à un système de référence, de préférence celui qui sera utilisé par la BD intégrée en le recalant à partir de points de référence du canevas de référence [Rouet 91].

Pour les bases de l'IGN sur Marne-la-Vallée, deux systèmes de référence sont utilisés, avec trois points de références [Direction technique de l'IGN 91]. La BD CARTO et GÉOROUTE sont en zone Lambert II étendue, et elles ont toutes les deux des points de référence différents. Par contre, la BD TOPO est en Lambert I. Ces trois bases ont pour point de référence le point inférieur gauche de leur emprise (l'origine du repère), ainsi toutes les coordonnées des géométries sont positives. Pour la BD intégrée, la projection en zone Lambert II étendue, a été choisie comme système de projection. Le point de référence a été défini afin de conserver les coordonnées positives (figure 40) :

pt_ref_BD_INTEGREE.x = min (pt_ref_BDT.x, pt_ref_BDC.x, pt_ref_G.x)
pt_ref_BD_INTEGREE.y = min (pt_ref_BDT.y, pt_ref_BDC.y, pt_ref_G.y)

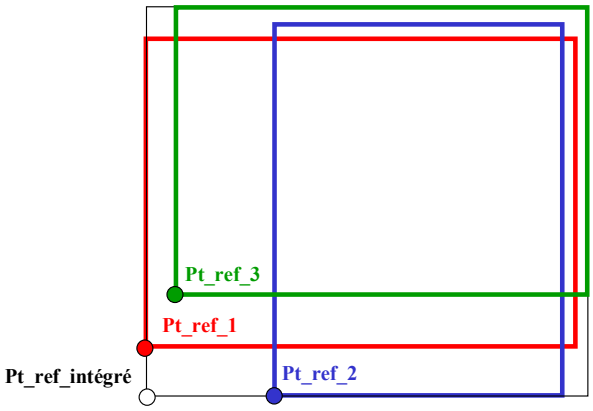



figure 40 : Point de référence de la BD intégrée

### *4.2.2 Enrichissement*

Comme nous l'avons exprimé auparavant (3.1.4.1), une traduction largement automatique est impossible sans l'enrichissement préalable des BDG. Dans ce but, il faut enrichir les bases par des méta-données, des mécanismes de conversion et des données implicites afin d'interpréter les schémas sans ambiguïté et d'obtenir la sémantique du monde réel qui ne peut pas être déduite à partir des schémas initiaux.

#### 4.2.2.1 Ajout de méta-données et de mécanismes de conversion

La concomitance de données issues de plusieurs bases pose des problèmes d'homogénéité des données (conflits de source, de précision, de résolution, …). Cependant, l'utilisateur de la BD intégrée doit connaître la « confiance » qu'il peut accorder au résultat de sa requête en fonction des données qui interviennent dans le traitement. Dans cet objectif, des **méta-données** doivent être ajoutées aux BDG à intégrer (source, date de dernière mise à jour, système de référence d'origine, point de référence, résolution, précision, exactitude, spécification de saisie, système de référence,…) [Stephan et al. 93]. De préférence, ces méta-données doivent être stockées selon le format standard défini [CEN/TC 287 95] ou [Federal Geographic Data Committee 94].

De plus, si nous désirons conserver dans la BD intégrée, les données dans leur format d'origine, des mécanismes de **traduction** (conversion de raster en vecteur et vice versa, changement de système de référence) doivent être insérés afin d'autoriser la manipulation conjointe de ces données incompatibles. La préservation de données selon leur format d'origine a l'avantage de ne pas dégrader les données par des traitements de conversion parfois inutiles et de définir le format commun en fonction du type de requête. Ainsi, pour manipuler deux jeux de données l'un en vecteur et l'autre en raster, le sens de la conversion dépend de la requête. Pour calculer l'intersection de deux surfaces, le format raster est sélectionné. Par contre, le format vecteur est utilisé pour rechercher une route dans une BDG vecteur qui ne passe pas dans une ville ayant une emprise stockée au format raster. Néanmoins, cette hétérogénéité alourdit les traitements.

Ces ajouts vont permettre de résoudre les conflits de mode de représentation et les conflits de type de positionnement et de signaler les conflits de méta-données de la géométrie.

#### 4.2.2.2 Extraction des données implicites de la base

Une des spécificités des BDG est la quantité d'information qui peut en être extrapolée [Grumbach et al. 96]. Pour intégrer des données en conflit de stockage de l'information (3.2.4.2), les phénomènes du monde réel représentés implicitement dans l'une des bases doivent être matérialisés, afin de parvenir au même niveau de compréhension.

Par exemple, la BD CARTO gère les embarcadères qui relient les tronçons de route et les tronçons de bac. Par contre, dans la BD TOPO, les instances de la classe TRONÇON_ROUTE ont pour extrémités des instances de la classe TRONÇON DE BAC (figure 41 (a) et (c)). Néanmoins, dans la BD TOPO, les embarcadères peuvent être déduits. Il a donc été décidé d'enrichir la BD TOPO en créant une classe EMBARCADERE (figure 41 (b)). Une instance de cette classe est obtenue pour chaque relation entre les instances des deux classes initiales, sa géométrie est l'intersection des deux géométries des instances en relation (figure 41 (d)).

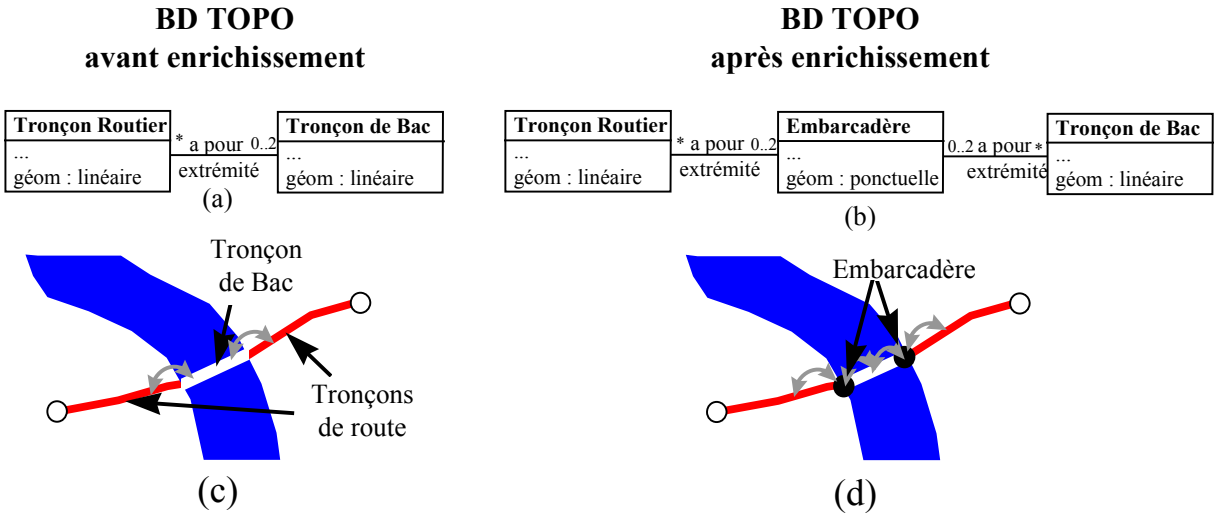



figure 41 : Enrichissement de la BD TOPO par ajout des embarcadères

Ces données implicites peuvent aussi être ajoutées à des classes existantes. Par exemple, dans GEOROUTE, le lieu où une route change de commune est représenté par une instance de la classe NOEUD_ROUTIER, ces instances ne sont pas présentes dans la BD TOPO. Toutefois, elles peuvent être déduites en superposant les limites administratives et les tronçons de routes. Pour matérialiser ces informations :

- une nouvelle valeur : « changement de commune » est ajoutée pour l'attribut énuméré *Type* de la classe NOEUD_ROUTIER,
- à chaque intersection entre un tronçon T et une limite L, un NOEUD_ROUTIER N dans la BD TOPO est créé (figure 42). Il prend pour l'attribut *Type* la valeur : « changement de commune ».
- Le tronçon T est scindé en deux tronçons T1 et T2 ayant chacun pour extrémité une des deux extrémités de T et N. Les attributs de T sont propagés ou partagés par T (ce problème sera abordé lors de la section suivante pour la règle 2).

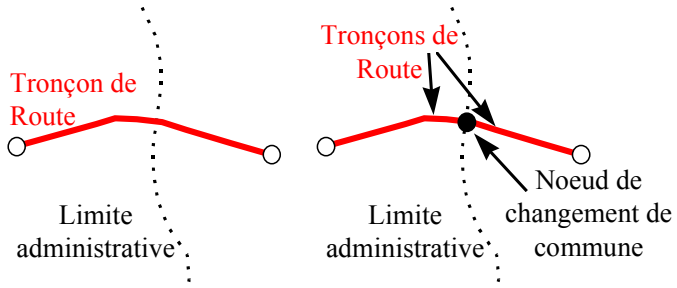



figure 42 : Enrichissement de la BD TOPO par ajout de Noeud routier de type « changement de communes »

L'extraction des données implicites et leur stockage dans la base vont permettre de résoudre une partie des conflits de stockage de l'information (3.2.4.2), l'autre partie sera traitée lors de l'application de la règle 1 de normalisation.

### *4.2.3 La normalisation*

La réduction des différences entre les schémas initiaux peut aussi être imposée par des **règles de normalisation** qui ont pour objectif de rendre :

- les classes plus homogènes,
- la modélisation du monde réel plus naturelle et ceci sans restrictions artificielles.

Pour les BDG de l'IGN, trois règles de normalisation sont imposées.

**Règle 1** : tous les phénomènes faisant partie de l'univers d'une classe doivent être des instances de cette classe sans restriction.

Cette règle a pour but de placer au même niveau tous les phénomènes du monde réel et de résoudre la plupart des conflits de stockage.

Elle est appliquée à la classe FRANCHISSEMENT de la BD CARTO V2. En effet, les instances de cette classe sont les lieux où plusieurs tronçons des réseaux routiers, ferrés ou hydrographiques s'intersectent sans qu'il y ait communication entre eux et avec les **restrictions** suivantes :

- le tronçon hydrographique est au dessous (le cas général),
- le tronçon de route est un chemin ou un sentier et le franchissement se fait à gué.

Cette restriction est contraire à la règle de normalisation 1. Ces deux contraintes sont supprimées. Ainsi chaque intersection entre un tronçon hydrographique et un autre tronçon est représentée par un objet de la classe FRANCHISSEMENT (figure 43). Cette instance présente une relation, *passe_sur* ou *passe_sous* avec des instances des classes TRONÇON_HYDROGRAPHIQUE, TRONÇON_ROUTE et TRONÇON_FERRE.

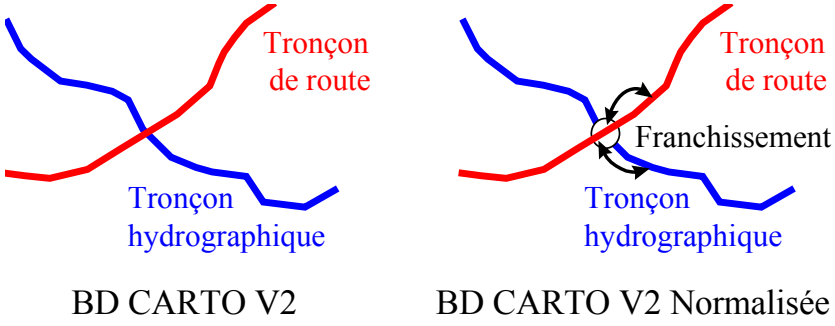



figure 43 : Normalisation des franchissements de la BD CARTO V2

Cette première règle de normalisation va permettre de résoudre l'autre partie des conflits de stockage de l'information (3.2.4.2).

**Règle 2 :** les instances d'une classe doivent représenter des phénomènes homogènes de même niveau de décomposition.

Cette règle a pour but de faciliter l'intégration en limitant les variantes d'intégration à l'intérieur d'une même classe et par conséquent de restreindre les conflits de fragmentations (3.2.3.3).

L'application de cette règle entraîne une modification de la classe TRONÇON_ROUTE de la BD TOPO car elle a un niveau de décomposition hétérogène. Effectivement, un tronçon de route du monde réel est représenté soit par une instance de TRONÇON_ROUTE soit par deux instances de cette classe (correspondant aux chaussés de la route du monde réel) et une instance de la classe SEPARATEUR. Les instances des classes SEPARATEUR et TRONÇON_ROUTE sont reliées indirectement par le partage partiel de leur géométrie. Pour que TRONÇON_ROUTE soit de même niveau de décomposition, les tronçons de chaussées séparées et les séparateurs sont transformés en tronçons de route.
Ainsi, les 4 tronçons de chaussées (T1, T2, T3, T4) et le séparateur S1 de la figure 44, sont normalisés en 3 tronçons de route :

- Ta est obtenu à partir de la géométrie de T1 et des attributs sémantiques de T1 et T3,
- Tb qui est obtenu à partir de l'intersection de la géométrie de T2 et T3 et des attributs sémantiques de T2 et T3,
- Tc qui est obtenu à partir de la géométrie de T4 et des attributs sémantiques de T2 et T4.

La solution inverse est aussi possible, elle aurait consisté à décomposer les tronçons de route en tronçons de chaussée.

La **valeur des attributs** des nouvelles instances est fonction des valeurs des anciens attributs.

Pour les attributs de **valeur homogène pour la composition** (attributs qui ont par nature la même valeur quelle que soit l'objet initial), leur valeur sera égale aux valeurs des objets initiaux. Ainsi, pour la classe TRONÇON_ROUTE l'attribut *état_chaussée* est de valeur homogène pour la composition (valeur identique à droite et à gauche du séparateur), la valeur de l'attribut du tronçon résultant est donc la valeur commune.

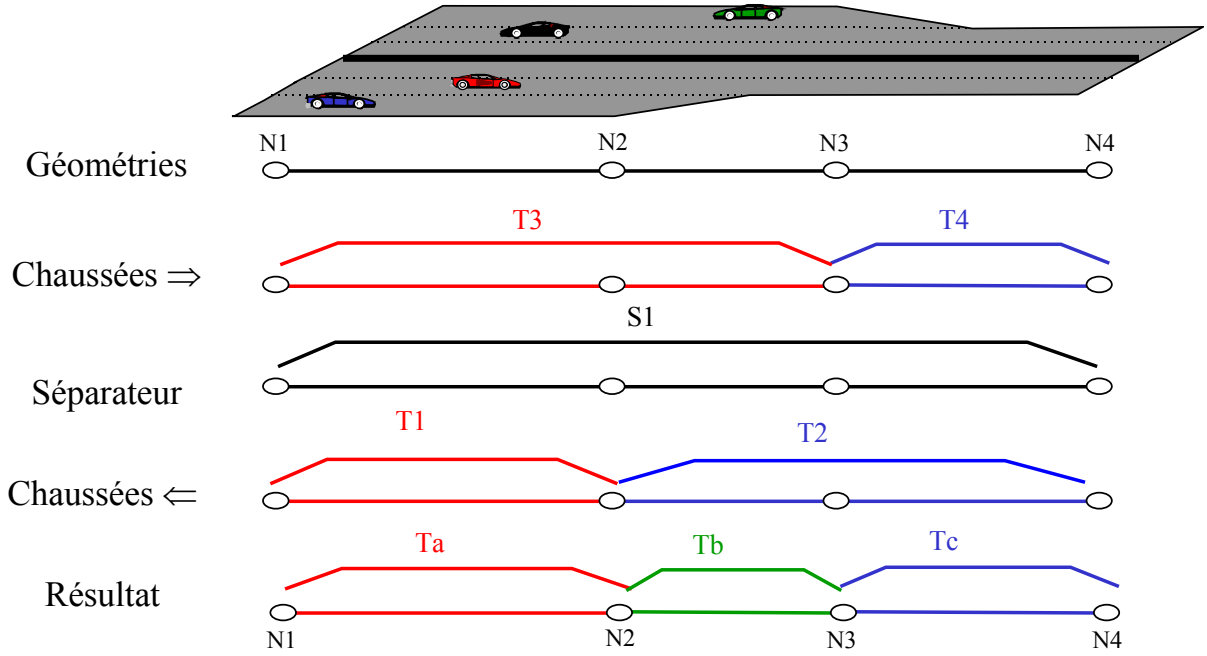



figure 44 : Normalisation des tronçons de route de la BD TOPO

Pour les attributs de **valeur hétérogène pour la composition**, des fonctions de transfert (somme, moyenne, …) doivent être définies. Malheureusement, ces fonctions risquent d'entraîner des pertes d'informations, lesquelles peuvent être évitées par la modification du type de l'attribut. Cette solution doit être complétée par une méthode renvoyant les données selon le format avant la modification. Par exemple, pour l'attributs *nb_voies,* de valeur hétérogène pour la composition, son type entier a été transformé en liste d'entier et une méthode donnant le nombre total de voies (somme des entiers de la liste) est ajoutée. Ainsi, la valeur de cet attribut est un singleton pour les tronçons sans séparateur et un couple (*nb_voies_sens_tronçon, nb_voies_sens_inverse)* pour les autres.

Le transfert des **instances des relations** (les liens) entre les anciens tronçons et les nouveaux est un peu plus complexe. En effet, un ancien lien peut être :

- partagé par toutes les nouvelles instances issues des instances portant ce lien,
- porté par une seule des nouvelles instances en fonction de la localisation. Des **règles de partage** doivent alors être définies. Elles s'appuient sur les liens homologues des objets initiaux, sur la topologie ou sur la géométrie.
  Pour les relations *passe_sur* et *passe_sous* des tronçons de route, la règle est : « un lien est créé pour une nouvelle instance, si tous les objets initiaux portent un lien du même type vers le même objet ». Ainsi, s'il existe un lien de la relation *passe_sur* entre T1 et un pont P et entre T3 et le même pont P, alors Ta récupère ce lien. Par contre, si T2 n'a pas de lien avec P, Tb n'aura pas de lien avec P.
  Des règles associées à la géométrie ou à la topologie peuvent aussi être définies, comme :
  « deux nouveaux objets (O1, O2) sont reliés par un lien de la relation R1
  si la géométrie de O1 est en relation topologique R2 avec la géométrie de O2 ».
  Pour la relation *a_pour_extremité*, de cette règle générique dérive la règle suivante : « une nouvelle instance de TRONÇON est en relation *a_pour_extremité* avec un objet d'une classe CARREFOUR si sa géométrie a pour extrémité la géométrie du carrefour. Ainsi, des liens Ta-N1, Ta-N2, Tb-N2, Tb-N3, Tc-N3 et Tc-N4 sont créés.

La dernière règle de normalisation est la suivante :

**Règle 3** : une classe ne doit pas avoir pour unique rôle de porter la géométrie d'une autre classe (classe d'objets géométriques).

Cette règle permet de résoudre certains conflits de définition de la géométrie. Par exemple, la classe COMMUNE de la BD TOPO est composée d'instances de la classe LIMITE_ADMINISTRATIVE qui représentent ses limites. Cette dernière porte uniquement un attribut sémantique qui peut être déduit des classes COMMUNE, ARRONDISSEMENT, DEPARTEMENT et REGION. La classe LIMITE_ADMINISTRATIVE est donc en opposition avec la règle 3. Pour normaliser la BD TOPO, une géométrie surfacique est calculée pour les objets de la classe COMMUNE à partir des géométries de la classe LIMITE_ADMINISTRATIVE et de la relation reliant ces deux classes, puis la classe LIMITE_ADMINISTRATIVE est détruite.

Une fois la règle 3 appliquée, les BDG ne comportent plus de classes d'objets géométriques.

### *4.2.4 Conclusion sur la pré-intégration de BDG*

La pré-intégration est une étape indispensable pour les BDG, elle résout :

- les conflits de modèle,
- les conflits de type de positionnement,
- les conflits de mode de représentation de la géométrie,
- les conflits de stockage de l'information.

Elle permet aussi de signaler les conflits de sources de données, les conflits d'abstraction de la troisième dimension, et les conflits de méta-données de la géométrie.

La pré-intégration facilite donc la suite du processus d'intégration :

- en fixant les fondements de la BD intégrée (le modèle commun, le système de référence, les mécanismes de conversion, les méta-données),
- en simplifiant la déclaration de correspondance, par l'homogénéisation des classes à intégrer et l'extraction et le stockage des phénomènes représentés implicitement,
- en réduisant le nombre de mises en conformité nécessaires lors de la phase d'intégration.

## 4.3 Extensions du langage de déclaration et de l'intégration des BDG

La phase de pré-intégration met en conformité les BDG. Cependant, des problèmes majeurs n'ont pas été traités. Ils sont divisés en trois parties.

- Les **conflits habituels** des BD classiques (conflits de structure, conflits de description simple) seront déclarés et résolus en employant les techniques d'intégration classiques déjà présentées.
- Les **conflits de modélisation de la topologie.** La différence de modèle topologique peut entraîner un découpage des géométries des objets différents (partage des primitives géométrique) et une absence d'une partie des relations topologiques dans la BDG ayant le modèle topologique le plus simple. L'utilisation de $GéO_2$ permet de faire cohabiter des données ayant des relations topologiques différentes. Cependant, il faut traiter, les problèmes de découpage (similaires aux conflits de fragmentation pour les objets géographiques) si on veut fusionner les géométries des instances homologues.
- Les autres conflits majeurs qui regroupent les **conflits spécifiques aux BDG** et les **conflits complexes** communs aux BD classiques et aux BDG. Pour ces conflits, les deux dernières phases (déclaration des correspondances et intégration) doivent être étendues afin

d'autoriser la déclaration des conflits spécifiques aux BDG et de disposer des techniques de résolution.

Ce chapitre va donc présenter ces extensions conflit par conflit (conflit de classification en 4.3.1, conflit de fragmentation en 4.3.2, conflit de critère de spécification en 4.3.3, conflit de description n-aire en 4.3.4, conflit de granularité en 4.3.5, conflits de description de la géométrie en 4.3.6). Mais auparavant, les extensions générales préliminaires vont être présentées (4.3.1).

## *4.3.1 Les extensions préliminaires*

Trois extensions préliminaires globales sont nécessaires :

- l'ajout de la notion de direction dans la syntaxe des déclarations de correspondances (4.3.1.1),
- la définition de stratégie d'intégration spécifique aux BDG (4.3.1.2),
- la définition d'une clause ou d'un mécanisme permettant de trouver dans les différentes BDG, les objets représentant les mêmes phénomènes du monde réel (4.3.1.3).

### 4.3.1.1 Ajout de la notion de direction

Certains éléments des BDG représentent des phénomènes de type **réseau orienté** (réseau routier, ferré, hydrographique, …). Pour ces éléments, les valeurs des attributs peuvent être fonction de la direction des arêtes. Celles-ci sont définies soit par des relations avec un sommet de départ et un sommet d'arrivée, soit par l'ordre des points intermédiaires formant l'arête. Par exemple, pour les tronçons routiers de la BD CARTO et de GEOROUTE, les attributs *nb_voies_sens_tronçon* et *nb_voies_sens_inverse* sont par construction fonction de la direction. Or, la direction d'une arête n'a aucune raison d'être similaire à la direction de l'arête en correspondance. Ainsi, pour l'exemple la valeur de *nb_voies_sens_tronçon* correspond soit à la valeur de *nb_voies_sens_tronçon* si la direction est semblable, soit à *nb_voies_sens_inverse* si la direction est opposée. Il faut donc prendre en compte la notion de direction dans les clauses AAC. Dans cet objectif, le prédicat **Direction (X,Y)** est ajouté au langage de définition. Il autorise l'égalité des attributs si deux objets linéaires orientés X et Y sont dans la même direction. L'AAC est du type :

$$\textbf{BD1.el1.att1} = _{\textbf{Direction(el1,el2)}}\ \textbf{BD2.el2.att2}$$

Pour l'exemple, les AAC suivantes sont obtenues :

$$\text{BDC.TRONÇON.nb\_voies\_sens\_tronçon} = _{\text{Direction (BDC.TRONÇON, GEOROUTE.TRONÇON)}} \text{G.TRONÇON.nb\_voies\_sens\_tronçon}$$

$$\text{BDC.TRONÇON.nb\_voies\_sens\_tronçon} = _{\neg\ \text{Direction (BDC.TRONÇON, GEOROUTE.TRONÇON)}} \text{G.TRONÇON.nb\_voies\_sens\_inverse}$$

### 4.3.1.2 Définition de stratégie d'intégration pour les BDG

Un grand nombre de techniques d'intégration très hétérogènes sont envisageables. Il est indispensable avant d'intégrer, de choisir une stratégie d'intégration afin de faciliter le choix de la technique d'intégration à appliquer en fonction du conflit et d'obtenir une base intégrée uniforme et conforme à notre objectif. Ce choix s'appuie sur un grand nombre de critères. Si nous considérons les 6 critères définis par Dupont [Dupont 95 b] comme indépendants, nous obtenons alors 64 ($2^6$) stratégies possibles. Ce qui reporte le problème du choix local des

techniques d'intégration à un choix global d'une stratégie d'intégration qui reste néanmoins complexe. Dans le cadre de l'intégration de BDG, le nombre de stratégies possibles a volontairement été réduit à 2 stratégies, très éloignées. En effet, pour constituer des BDG centralisées intégrées, deux approches semblent se dégager : la stratégie mono-représentation, ou fusion, et la stratégie multi-représentation ou préservation.

#### *4.3.1.2.1 Stratégie mono-représentation ou fusion*

La stratégie mono-représentation (ou fusion) a pour objectif de produire une **unique représentation** du monde réel, s'appuyant sur les informations les plus précises dans chacune des bases initiales. Cette stratégie répond donc aux critères suivants :

- de **non conservation** : les informations les moins précises sont supprimées,
- de **précision** : la précision initiale est conservée,
- de **complétude** : toutes les redondances sont supprimées,
- de **non réversibilité** : les informations initiales ne peuvent pas être déduites de la BD intégrée,
- **d'unification** : les éléments créés dans la BD intégrée, regroupent toutes les occurrences des BD initiales.
- de **liberté** : les techniques employées doivent être sans condition préalable.

Les techniques utilisées pour cette stratégie devront répondre aux six critères énoncés ci-dessus. De plus, une préférence est accordée aux techniques produisant le schéma optimum.

#### *4.3.1.2.2 Stratégie multi-représentation ou préservation*

L'objectif de la stratégie multi-représentation (ou préservation) est radicalement différent. Les différentes représentations du monde réel ne sont pas fusionnées, mais les éléments représentant les mêmes phénomènes du monde réel sont reliés entre eux. Cette stratégie répond aux critères suivants :

- de **conservation** : toutes les informations sont conservées ou peuvent être calculées,
- de **précision** : la précision initiale est conservée,
- de **non complétude** : toutes les redondances ne sont pas supprimées, la même information est conservée à différents niveaux de détail.
- de **réversibilité** : les informations initiales peuvent être déduites de la BD intégrée, les informations de la BD intégrée peuvent être réaffectées sur les BD initiales.
- **d'unification** : les éléments créés dans la BD intégrée regroupent toutes les occurrences des BD initiales.
- de **liberté** : les techniques utilisées doivent être sans condition préalable.

Les techniques utilisées pour cette stratégie doivent répondre aux six critères précédents. En outre, les techniques utilisées doivent permettre une intégration des données la plus simple possible.

#### *4.3.1.2.3 Comparaison de ces deux stratégies*

Ces deux stratégies gèrent de manière similaire les éléments prenant les mêmes valeurs pour le même phénomène du monde réel. Par contre, pour les éléments ayant la même sémantique, mais prenant des valeurs différentes pour le même phénomène du monde réel, le traitement n'est pas identique. La stratégie mono-représentation choisit la représentation la plus détaillée alors que la stratégie multi-représentation va relier les différentes représentations du même phénomène.

Une stratégie sera sélectionnée en fonction de l'objectif de la BD intégrée et des BDG à intégrer. Pour des BDG initiales « **proches** » (échelle, contexte,…) la stratégie mono-représentation est préférable car la concomitance des deux représentations alourdit la BDG intégrée sans l'enrichir. Cette notion de proximité est détaillée dans [Brugger 94] [Brugger 95] elle est appelé **cohérence**. Par contre, pour des BD « éloignées », une stratégie multi-représentation doit être employée.

#### *4.3.1.2.4 La stratégie d'intégration définie pour les BD de l'IGN*

Pour intégrer les trois BDG de l'IGN, il a été décidé de procéder en deux temps et d'employer les deux stratégies successivement. Dans un premier temps, une stratégie mono-représentation est appliquée pour intégrer la **BD TOPO** et **GEOROUTE**. Le résultat est appelé la **BD intégrée 1 (BDI 1)**. Cette stratégie est choisie, car ces deux BD ont une échelle comparable et sont complémentaires. En effet, la géométrie de la BD TOPO est plus précise, alors que les attributs sémantiques de GEOROUTE sont plus détaillés pour le thème routier.

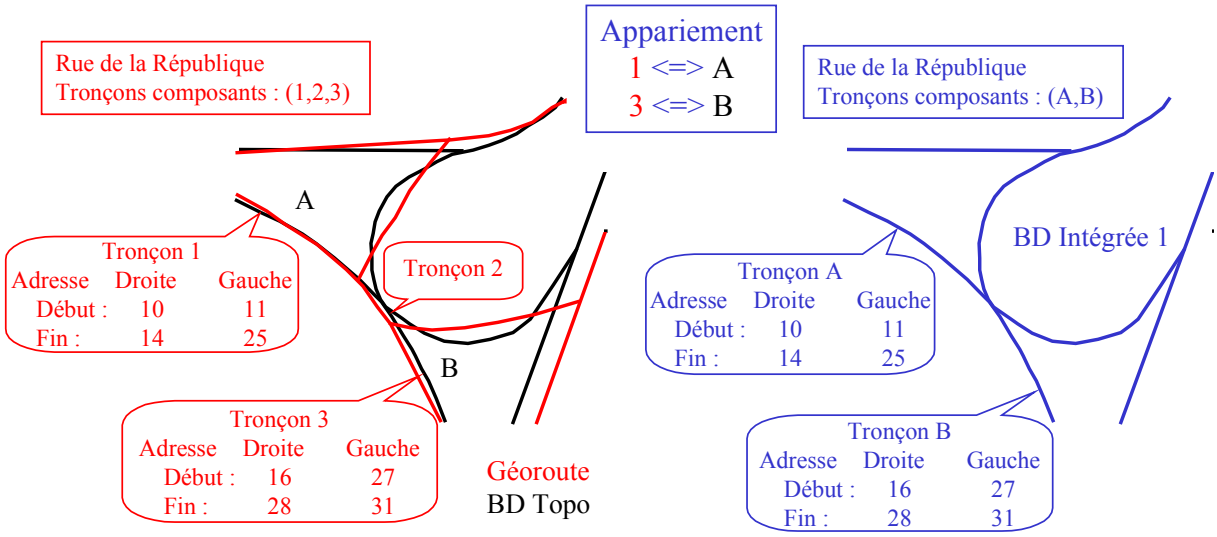



figure 45 : Intégration de la BD TOPO et de GEOROUTE

La figure 45 montre l'intérêt d'une telle stratégie. Nous remarquons que la BD TOPO (en noir) possède une géométrie plus précise, le rendu du virage est plus lisse. Cependant, GEOROUTE (en rouge) bénéficie d'informations telles les adresses, les noms des rues que ne possède pas la BD TOPO. La BDI 1 résultante profite alors de la complémentarité des deux bases. En outre, elle autorise des applications nouvelles, pour lesquelles les BDG initiales prises indépendamment, sont insuffisantes. Ainsi, l'application **plans de villes** qui nécessite une géométrie précise afin de produire des cartes papiers au 1 : 10 000 et des informations sémantiques détaillées telles que le nom des rues et les adresses des débuts et fins de tronçons

profiterait de l'intégration de ces deux BDG (actuellement, cette application utilise uniquement le réseau routier de GEOROUTE).

Dans un deuxième temps, la BDI 1 et la BD CARTO sont intégrées. Compte tenu de la différence d'échelle, une stratégie multi-représentation est utilisée. Le résultat de cette intégration est appelé **BD intégrée 2 (BDI 2)**.

La BD issue de l'intégration des 3 principales BDG de l'IGN dispose de deux représentations, une détaillée provenant de BDI 1 et une moins détaillée issue de la BD CARTO et de l'ensemble de l'information sémantique.

#### 4.3.1.3 Déclaration des clauses « Appariement Géométrique des Données » ou définition d'un processus d'appariement

Pour rendre l'intégration opérationnelle, il faut pouvoir identifier dans les BDG initiales, l'ensemble des objets représentant le même phénomène du monde réel. Effectivement, toute l'information des bases initiales est disponible dans la BD intégrée si, et seulement si, ces objets sont intégrés.

Pour ce faire, chaque ACI doit comprendre une clause de spécification de la correspondance entre les instances. Le plus souvent pour les BD classiques, il existe au moins un identifiant qui peut être employé pour cette fonction : une clause « Avec Identifiants Correspondants » (AIC) est donc déclarée. Or, dans les BDG, ces identifiants communs sont rarement présents. Toutefois, par leur localisation, les BDG offrent une alternative pour identifier les objets [Laurini et Thompson 92]. La localisation permet d'avancer que deux instances se correspondent si elles sont approximativement situées au même endroit. La clause AIC est donc relâchée au profit d'une clause **appariement géométrique des données (AGD)** qui spécifie le prédicat d'appariement entre les instances. Ce prédicat peut être une correspondance entre :

- des attributs identifiants,
- des géométries,
- des relations topologiques.

Des correspondances complexes peuvent aussi être définies. Elles utilisent des fonctions standard ou des méthodes ad-hoc qui prennent comme paramètres les trois types de données. La fonction standard la plus employé est G **INSIDE**(S) qui vérifie si la géométrie G (point, ligne ou surface) est à l'intérieur d'une surface. Cette fonction peut être précédée par la fonction **BUFFER**(G, d) qui transforme une géométrie G en une surface S. Cette surface est définie telle que pour tout point P inclus dans S, il existe un point P2 de G à une distance inférieure ou égale à d. Le prédicat G1 **INSIDE**(**BUFFER**(G2, d)) peut être ainsi utilisé. Par exemple, pour les correspondances entre les instances de la classe NOEUD de la BD CARTO et les instances des classes NOEUD et TRONÇON de la BD TOPO (tableau 6), l'AGD est présenté dans le tableau 6 :

**ACI** : BDC.NOEUD ⊆ SET([1:N] BDT.NOEUD,[0:N]BDT.TRONÇON)
**AGD** : SET([1:N] BDT.NOEUD) = { nt / nt BDT.NOEUD ∧
nt INSIDE (BUFFER(BDC.NOEUD,2 × BDC.exactitude)}
SET([0:N] BDT.TRONÇON) = { tt/ tt ∈ BDT.TRONÇON ∧
∃ nt ∈SET([1:N] BDT.NOEUD( debut(tt,nt)) ∧
∃ nt ∈SET([1:N] BDT.NOEUD ( fin (tt,nt)) }

tableau 6 : Exemple de clause Appariement Géographique des Données (AGD)

Cette clause AGD, permet de lier les instances de la classe NOEUD de la BD CARTO aux instances de la classe NOEUD à une distance inférieure à deux fois l'exactitude de la BD CARTO, et aux instances de la classe TRONÇON de la BD TOPO ayant pour noeud initial et noeud final des noeuds appariés avec le noeud de la BD CARTO.

Cependant, les données des BDG présentent le plus souvent des différences importantes et l'appariement des données doit alors combiner un grand nombre de méthodes et s'appuyer sur des données déjà appariées. Pour l'exemple du tableau 6, cette clause AGD n'est pas suffisante. En effet, elle permet de sélectionner l'ensemble des tronçons et des noeuds s'appariant avec le noeud de la BD CARTO. Mais, elle risque aussi de relier le noeud de la BD CARTO avec des éléments « **parasites** » (figure 75 et figure 76), c'est-à-dire des éléments qui ne devraient pas être appariés avec ce noeud mais qui sont à une distance proche, il faudrait alors ajouter des **filtres**. Dans ce cas, il est difficile de déclarer de manière concise la clause AGD. Qui plus est, la **distance de la zone tampon** est difficile à définir. Elle devrait s'appuyer sur une erreur maximum. Cette méta-donnée n'est malheureusement pas présente dans les BDG. On dispose, le plus souvent, uniquement d'une erreur moyenne (l'exactitude). Les clauses AGD des différentes ACI doivent donc être le plus souvent remplacées par un processus d'appariement global. Celui-ci sera décrit dans le partie 5.

Pour résumer, si la spécification de la correspondance est relativement simple, une clause AGD est comprise dans chaque ACI. Par contre, si la spécification s'avère complexe (appariement interdépendant, filtrage, …) l'ensemble des clauses AGD est remplacé par un processus d'appariement. Pour les BDG de l'IGN, nous sommes dans le deuxième cas.

### *4.3.1 Résolution des conflits de classification*

Les conflits de classification sont dus à des abstractions de classification différentes dans les BD (3.2.3.1).

#### 4.3.1.1 Expression des conflits de classification

Le langage de déclaration de correspondance classique permet de signaler des conflits de classification entre deux classes des deux schémas en utilisant une **relation ensembliste** différente de l'équivalence.

**BD1.ELEMENT1 <RE> BD1.ELEMENT1 (RE ≠ ≡)**

Des assertions plus génériques qui relient plusieurs éléments avec plusieurs éléments ou ayant des relations n-m au niveau de leurs instances (conflits de fragmentations) doivent aussi être exprimées. Dupont ([Dupont 95 b] [Dupont 94]) a proposé des extensions du langage pour exprimer ces conflits, mais elles sont trop complexes. Nous avons donc proposé une autre syntaxe pour inclure la déclaration de ces deux types de conflits dans l'entête de l'ACI.

Pour les conflits de classification n-m, la syntaxe est étendue par ajout de l'ensemble des éléments (classes, relations) entre parenthèses de chaque base.

Ainsi, pour décrire des correspondances entre plusieurs éléments et un élément, l'ACI obtenue est du type :

**BD1.ELEMENT1 <RE> BD2.(ELEMENT21, ELEMENT22,…)**

De même, pour décrire des correspondances entre plusieurs éléments et plusieurs éléments, l'ACI est du type :

**BD1.(ELEMENT11, ELEMENT12,…) <RE> BD2.(ELEMENT21, ELEMENT22,…)**

Pour l'exemple de la figure 27 (page 3), l'ACI entre les classes de la base Eau et de la base Forêt est la suivante :

BD_EAU.(ZONE_HUMIDE, ZONE_BOISEE) ≡ BD_FORET. (ZONE_HUMIDE, ZONE_BOISEE)

#### 4.3.1.2 Intégration de classes en conflit de classification

##### *4.3.1.2.1 Intégration de classes en conflit de classification au niveau de la relation ensembliste*

Pour résoudre les conflits de classification au niveau de la relation ensembliste (BD1.ELEMENT1 <RE> BD2.ELEMENT1 (RE ≠ ≡)), plusieurs techniques (annexe 7.3.1) sont envisageables :

- la **fusion**, est la plus simple. Elle n'est soumise à aucune condition préalable mais a pour inconvénient de produire une classe avec des attributs optionnels (attributs propres à une des classes d'origine).
- La **sous-classe** ou réciproquement la **sur-classe** sont plus précises mais produisent des schémas plus compliqués. Elles peuvent être employées pour les relations ensemblistes du type $\supset, \supseteq, \subset, \subseteq$
- La **généralisation** ou des techniques plus évoluées sont envisageables pour la relation ensembliste d'intersection ($\cap$).

Afin de produire une méthode générique sans condition préalable, nous avons retenu la technique de fusion pour les deux stratégies.

##### *4.3.1.2.2 Intégration de classes en conflit de classification entre une classe et plusieurs classes*

Plus généralement, dans les conflits de classification interviennent plusieurs classes de chaque base à intégrer. Pour les conflits de classification entre une classe et plusieurs classes, deux techniques d'intégration sont sélectionnées : la partition et la fusion.

La **partition** produit, dans le schéma intégré, les classes :

- CLASSE11-CLASSE2 ayant pour instances les objets issus d'un couple d'objet de CLASSE11 et de CLASSE2 représentant le même phénomène du monde réel.
- CLASSE12-CLASSE2 ayant pour instances les objets issus d'un couple d'objet de CLASSE12 et de CLASSE2 représentant le même phénomène du monde réel.
- CLASSE1i-CLASSE2 …

Si la relation ensembliste est différentes de l'équivalence, cette intégration produit d'autres classes :

- CLASSE2-, pour les relations $\subset$, $\subseteq$ ou $\cap$, qui a pour instances, les objets de la CLASSE2 n'ayant pas de correspondant dans les CLASSE1i.
- CLASSE1i-, pour les relations, $\supseteq$, $\supset$ ou $\cap$, qui ont pour instances, les objets de la CLASSE1i n'ayant pas de correspondant dans les CLASSE2.

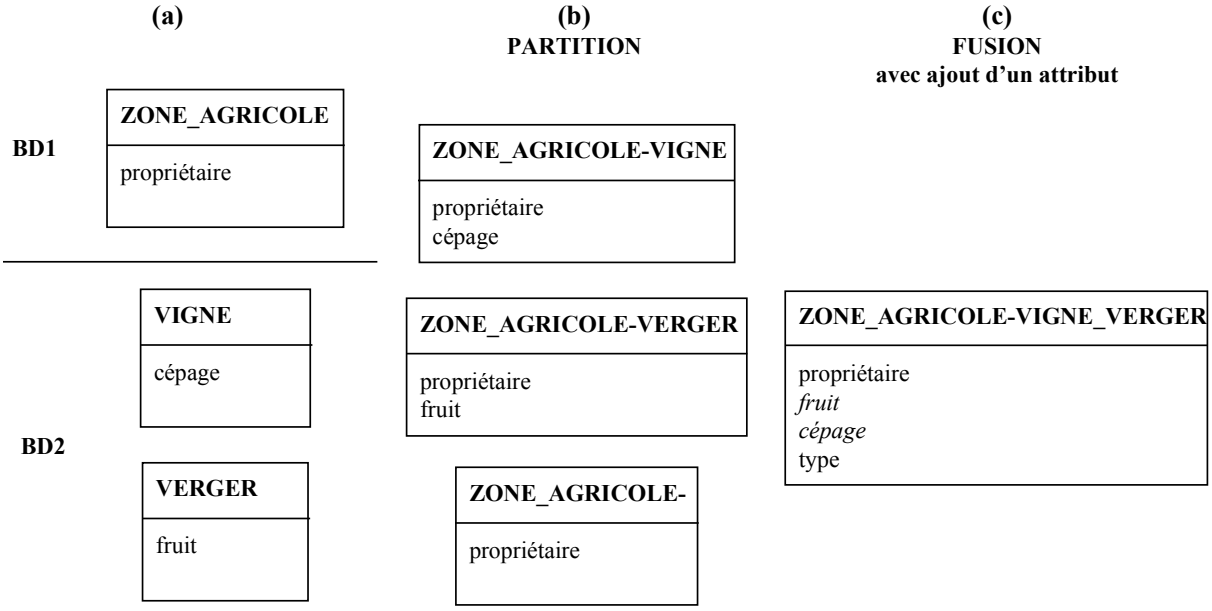



figure 46 : Exemple d'intégration de classe en conflit de classification 1-n

Ainsi, l'intégration par partition de BD1.ZONE_AGRICOLE $\supset$ BD2.(VIGNE, VERGER) va créer 3 classes dans la BD intégrée (figure 46 b) :

- ZONE_AGRICOLE-VIGNE,
- ZONE_AGRICOLE-VERGER,
- ZONE_AGRICOLE-.

La deuxième technique proposée ; la **fusion avec ajout d'un attribut,** reporte la classification au niveau d'un nouvel attribut énuméré qui prend les valeurs CLASSE11-CLASSE2, CLASSE12-CLASSE2,…, voire CLASSE2- ou CLASSE1i- si nécessaire. Les attributs et les relations propres à une classe ayant des instances sans correspondant deviennent optionnels.

L'intégration par fusion du même exemple va créer une classe avec 4 attributs dans la BD intégrée (figure 46 c) :

- propriétaire qui est obligatoire,
- fruit qui est optionnel,
- cépage qui est optionnel,
- type qui est le nouvel attribut de classification. Il a pour valeur possible vigne, verger ou autre.

#### *4.3.1.2.3 Intégration de classes en conflit de classification entre plusieurs classes et plusieurs classes*

Le conflit de classification peut être aussi entre plusieurs classes et plusieurs classes. Dans ce cas, les deux mêmes techniques d'intégration sont possibles. La **partition** générera des classes :

- CLASSE1i-CLASSE2j ayant pour instances, les objets issus d'un couple d'objets de CLASSE1i et de CLASSE2j représentant le même phénomène du monde réel,
- des classes CLASSE1i- si la relation ensembliste est $\supset$, $\supseteq$ ou $\cap$,
- des classes CLASSE2i- si la relation ensembliste est $\subset$ $\subseteq$ ou $\cap$.

Par exemple, pour intégrer la base sur l'eau et la base sur la forêt (figure 27 page 3), l'ACI est :

BD_EAU.(ZONE_HUMIDE, ZONE_BOISEE) ≡ BD_FORET. (ZONE_HUMIDE, ZONE_BOISEE)

La partition fournit les trois classes :

- ZONE_HUMIDE-ZONE_HUMIDE qui a pour instances les objets issus de l'intégration d'un objet de la classe BD_EAU.ZONE_HUMIDE avec un objet de la classe BD_FORET.ZONE_HUMIDE,
- ZONE_HUMIDE-ZONE_BOISEE qui a pour instances, les objets issus de l'intégration d'un objet de la classe BD_EAU.ZONE_HUMIDE avec un objet de la classe BD_FORET.ZONE_BOISEE,
- ZONE_BOISEE-ZONE_BOISEE qui a pour instances les objets issus de l'intégration d'un objet de la classe BD_EAU.BOISEE avec un objet de la classe BD_FORET.ZONE_BOISEE.

La classe ZONE_BOISEE-ZONE_HUMIDE sera supprimée, car elle n'a pas d'instance.

La **fusion avec ajout d'un attribut** est similaire. Le nouvel attribut de type énuméré a pour valeur possible les noms des classes qui auraient été déterminées par la partition.

Pour intégrer les classes en conflit de classification, les techniques choisies sont volontairement simples. Le mécanisme de généralisation entre autre, n'est pas utilisé afin de ne pas créer de classe générique inutile. Par contre, la technique de partition génère un grand nombre de classes. Il faut donc laisser à l'administrateur la possibilité de définir des **post traitements** de modification de schémas [Scherrer et al. 93] [Scholl and Tresch 93], permettant de définir les classes génériques utiles et de regrouper des classes issues de la partition. De même, pour la technique de fusion, certaines valeurs du nouvel attribut doivent pouvoir être regroupées.

Pour intégrer les BDG de l'IGN, nous avons retenu la technique de fusion avec ajout d'un attribut pour les deux stratégies, car elle n'augmente pas le nombre de classes dans la BDG intégrée.

### *4.3.2 Résolution des conflits de fragmentation*

Il existe un **conflit de fragmentation** [Dupont 95a] quand un objet dans une base correspond à plusieurs objets dans l'autre base (3.2.3.3).

#### 4.3.2.1 Expression des conflits de fragmentation

Pour la déclaration des **conflits de fragmentation** la description des ACI est surchargée en rendant possible la déclaration d'**éléments virtuels.** Les instances de ces éléments virtuels

regroupent des instances d'un élément initial en correspondance avec la même instance de l'autre base. Un élément virtuel est défini en précédant les éléments par :

- l'expression SET,
- un couple d'entiers qui représentent la cardinalité minimum et la cardinalité maximum de l'ensemble formé par les fragments.

Une ACI décrivant un conflit de fragmentation est du type :

**BD1.ELEMENT1 <RE> BD2.SET([1:N] ELEMENT2)**

SET([1:N] ELEMENT2) représente un élément virtuel (classe virtuelle, relation virtuelle)

Pour des conflits de fragmentation n-m, l'ACI est du type :

**BD1.SET([1:N] ELEMENT1) <RE> BD2.SET([1:N] ELEMENT2)**

Pour l'exemple sur les conflits de fragmentation, entre les tronçons de la BD CARTO et ceux de la BD TOPO (un tronçon de la BD CARTO correspond à plusieurs tronçons de la BD TOPO voir 3.2.3.3) l'ACI suivante est obtenue :

BDC.TRONÇON $\subseteq$ G.SET([1:N] TRONÇON)

Les instances des éléments virtuels sont construites à partir des instances fragments. Pour construire les instances virtuelles, des **contraintes** peuvent être spécifiées. Dans l'exemple précédant, les contraintes suivantes semblent raisonnables :

- les tronçons fragments d'une même instance virtuelle doivent appartenir à la même route,
- les tronçons fragments d'une même instance virtuelle doivent être contigus.

Ces contraintes peuvent être exprimées comme des prédicats d'un langage de requête au niveau de la déclaration de l'AGD (si cette clause est conservée) ou sinon dans le **processus d'appariement**.

Dans une même ACI, les conflits de fragmentation et les conflits de classification peuvent coexister. Nous obtenons alors des ACI du type :

**BD1.SET([i:j] ELEMENT11, [i:j] ELEMENT12,…)**
**<RE>**
**BD2.SET( [i:j] ELEMENT21, [i:j] ELEMENT22,…)**

Pour l'exemple du conflit de décomposition (figure 31), l'ACI suivante est obtenue :

BD1.TRONÇON_ROUTE
≡
BD2.SET( [1:2] TRONÇON_CHAUSSEE, [0:1] SEPARATEUR)

#### 4.3.2.2 Intégration de classes en conflit de fragmentation

Pour intégrer des classes en conflit de fragmentation, la technique employée dépend de la cardinalité du conflit et de la stratégie.

##### *4.3.2.2.1 Intégration de classes en conflit de fragmentation 1-n pour une stratégie mono-représentation*

Pour intégrer des classes en conflit de fragmentation 1-n (X <RE> SET([i:j] Y)) avec une stratégie mono-représentation, il est nécessaire de conserver l'information la plus fragmentée (figure 47), afin de maintenir l'information la plus précise possible.

Une instance y+ de la classe Y+ du schéma intégré est obtenue à partir d'une instance y de Y. Elle reçoit les valeurs de y et de x (l'instance de X en correspondance avec y).

Pour les relations ensembliste ⊃ et ⊇, des phénomènes du monde réel représentés dans X, ne sont pas représentés dans Y. Il faut alors intégrer ces instances de X dans la BD intégrée selon le format de Y. Cette transformation est rarement nécessaire, car les classes les plus fragmentées sont souvent les classes représentant le plus de phénomènes du monde réel.

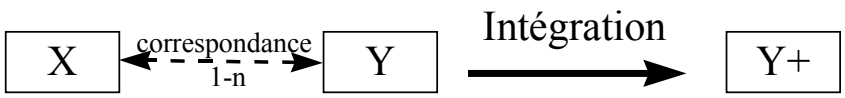



figure 47 : Intégration de classe en conflit de fragmentation 1-n
avec une stratégie mono-représentation

#### *4.3.2.2.2 Intégration de classes en conflit de fragmentation 1-n pour une stratégie multi-représentation*

Pour intégrer des classes en conflit de fragmentation 1-n (X <RE> SET([i:j] Y)) avec une stratégie multi-représentation (figure 48), une **classe virtuelle** (Y') ayant pour instances virtuelles des objets formés par l'agrégation des objets correspondant au même objet de X est créée. Ainsi, à chaque objet de la classe Y' correspond un objet de la classe X. X et Y' sont donc en correspondance 1-1 au niveau de leurs instances et peuvent être intégrées par les méthodes classiques. La méthode de fusion est choisie pour intégrer X et Y' du fait de sa simplicité. Le schéma intégré inclut la classe Y, la classe X-Y' résultant de la fusion de X avec Y' et une relation de composition de Y vers X-Y'.

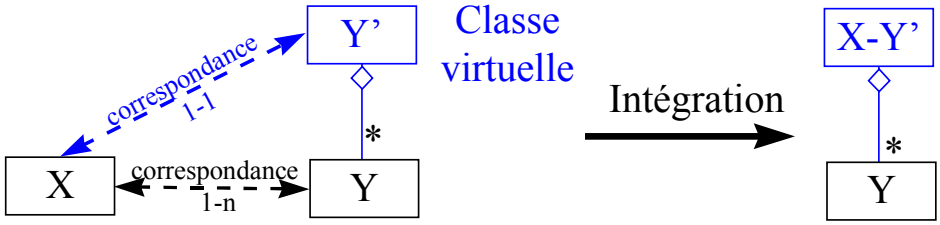



figure 48 : Intégration de classes en conflit de fragmentation 1-n
avec une stratégie multi-représentation

Les classes TRONÇON de la BD TOPO et TRONÇON de la BD CARTO sont en conflit de fragmentation 1-n. Le schéma intégré inclut donc :

- une classe TRONÇON_DETAILLE,
- une classe TRONÇON_GENERALISE issue de la fusion de la classe virtuelle ENSEMBLE_TRONÇON_BDT et TRONÇON de la BD CARTO,
- une relation de composition entre ces deux classes.

#### *4.3.2.2.3 Intégration de classes en conflit de fragmentation n-m pour une stratégie mono-représentation*

Pour intégrer des classes en conflit de fragmentation n-m (SET([i:j] X) <RE> SET([i':j'] Y)) avec une stratégie mono-représentation, les objets des deux bases doivent être divisés afin d'obtenir des relations 1-1.

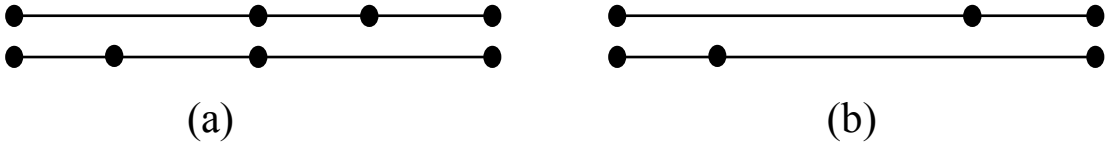



figure 49 : Exemple de conflit de fragmentations n-m

Si le conflit de fragmentation n-m est en réalité un conflit de fragmentation **1-n alterné**, c'est-à-dire qu'au niveau des instances, des sous classes peuvent être définies telles que pour chaque couple de sous classe il existe toujours soit des relations 1-n, soit des relations n-1 (figure 49 a), cette intégration est difficile mais envisageable (4.3.6.2.1.1).

Par contre, si nous sommes en présence d'un véritable conflit n-m (figure 49 b), il faudra régler tous les problèmes liés à la désagrégation de valeurs, aussi appelée interpolation [Flowerdew et Openshaw 87] [Weber 94], ce qui risque de dégrader les données et la fiabilité des valeurs des attributs. Ainsi, [Weber 94] pour comparer les données des recensements de 1990 et 1982 pour la population de Strasbourg selon des découpages en quartier incompatibles (48 quartiers en 1990 et 81 quartiers en 1982), a utilisé une image satellite SPOT 2 pour identifier les zones bâties et interpoler la variable *population* en fonction de la variable *zone bâtie*. Cette technique est plus fiable qu'une simple interpolation en fonction des rapports de surface commune, néanmoins elle dégrade la qualité des données, car cette corrélation n'est pas parfaite.

#### *4.3.2.2.4 Intégration de classes en conflit de fragmentation n-m pour une stratégie multi-représentation*

Pour intégrer des classes en conflit de fragmentation n-m (SET([i:j] X) <RE> SET([i':j'] Y)) avec une stratégie mono-représentation, des classes virtuelles X' et Y', sont employées. Leurs instances regroupent des instances des classes initiales et sont en correspondance 1-1.

Le conflit de fragmentation est alors traité comme une correspondance simple au niveau des classes virtuelles (X',Y'). Elles peuvent donc être intégrées par les mécanismes classiques d'intégration. Dans notre cas, le mécanisme de fusion est utilisé. Le schéma intégré inclut (figure 50) :

- la classe X,
- la classe Y,
- la classe X'-Y' résultant de la fusion de X' avec Y',
- une relation de composition entre X et X'-Y'
- une relation de composition entre Y et X'-Y'

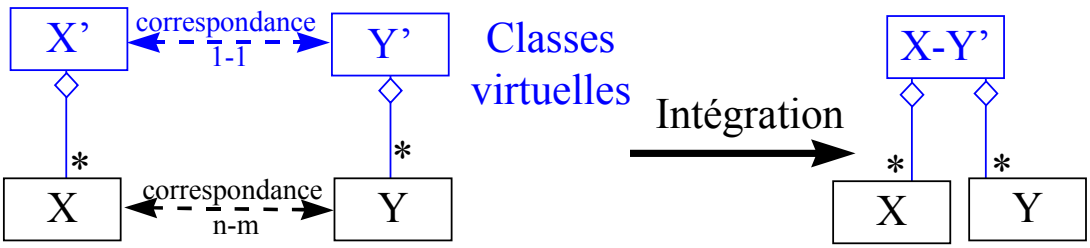



figure 50 : Intégration de classe en conflit de fragmentation n-m

La correspondance entre la classe PONT de la BD TOPO et la classe FRANCHISSEMENT de la BD CARTO est de ce type (un pont passant au-dessus de deux tronçons correspond à deux franchissements réciproquement deux ponts superposés correspondent à un

franchissement). Le schéma intégré inclut une classe PONT, une classe FRANCHISSEMENT, une classe ENSEMBLE_PONT_FRANCHISSEMENT résultant de la fusion des classes virtuelles ENSEMBLE_PONT et ENSEMBLE_FRANCHISSEMENT et les relations de composition.

#### 4.3.2.3 Intégration de classes en conflit de classification et de fragmentation

Les conflits de classification et de fragmentation peuvent aussi être combinés. Les déclarations de correspondance sont du type :

**BD1.SET([i11,j11] CLASSE11, [i12,j12] CLASSE12,…)**
**<RE>**
**BD2. SET([i21,j21] CLASSE21, [i22,j22] CLASSE22,…)**

Dans ce cas, il faut combiner la technique résolvant les conflits de classification avec la technique résolvant les conflits de fragmentation. Pour une **stratégie mono-représentation**, la combinaison de ces deux conflits ne demande pas d'extension. Par contre, pour une **stratégie multi-représentation**, il est nécessaire d'illustrer la technique résultant de la combinaison par un exemple portant sur la correspondance entre les classes NOEUD de la BD CARTO et les classes TRONÇON et NOEUD de GEOROUTE (figure 64 page 3).

L'ACI est :

BDC. NOEUD ⊂ BDT.SET([1,n] NOEUD, [0,m] TRONÇON)

Elles regroupent ces deux conflits.

Le schéma issu de l'intégration (figure 51) inclut :

- les classes :
  - BDT.NOEUD,
  - BDT.TRONÇON,
  - NOEUD_GENERALISE issues de l'intégration de la classe NOEUD de la BD CARTO et de la classe virtuelle ENSEMBLE_TRONÇON_NOEUD de la BD TOPO,
- les relations de composition :
  - entre NOEUD_GENERALISE et BDT.NOEUD,
  - entre NOEUD_GENERALISE et BDT.TRONÇON.

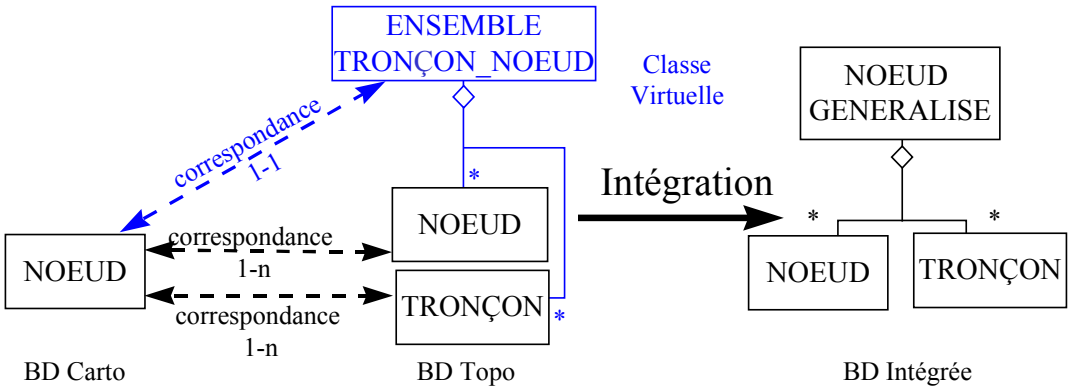



figure 51 : Intégration des classes NOEUD et TRONÇON de la BD TOPO en conflit de classification 1-n et de fragmentation 1-n avec le NOEUD de la BD CARTO

### *4.3.3 Résolution des conflits de critères de spécification*

Un conflit de critères de spécification est présent quand les critères de spécification (sélection ou décomposition) sont différents (3.2.3.2).

#### 4.3.3.1 Expression des conflits de critères de spécification

Pour inclure la déclaration de ces conflits, l'entête de l'ACI est surchargée.

##### *4.3.3.1.1 Déclaration des conflits de critères de sélection*

Pour les **critères de sélection**, deux solutions sont envisageables. La première consiste à déclarer le critère de sélection par une relation ensembliste différente de l'équivalence. Cette solution n'est pas satisfaisante car toute l'information disponible n'est pas présente dans la déclaration. La deuxième solution, plus précise, exprime le critère de sélection dans une sélection. L'ACI est alors du type :

**BD1.ELEMENT1 ≡ SELECTION (critère de sélection) BD2.ELEMENT2**

Par exemple, pour stipuler qu'une impasse de GEOROUTE de longueur inférieure à 50 mètres n'a pas de correspondance dans la BD TOPO, l'ACI est la suivante :

$$\text{BDT.TRONÇON} \equiv \underset{\text{G.TRONÇON}}{\text{SELECTION}} (\neg (\text{Impasse(G.TRONÇON)} \wedge \text{L(G.TRONÇON)} < 50\ \text{m}))$$

« Impasse » est une fonction qui teste si le tronçon est une impasse (un des noeuds extrémités est de degré[11] 1) et L est une fonction qui renvoie la longueur du tronçon.

#### 4.3.3.2 Déclaration des conflits de critères de décomposition

Les **critères de décomposition** sont plus complexes à exprimer. En effet, une ACI devant inclure un critère de décomposition doit déclarer qu'une instance d'un élément (une classe, une relation) correspond soit à une instance de l'élément en correspondance soit à un ensemble d'instances d'éléments en correspondance. Ces critères sont déclarés au niveau de la relation ensembliste sous la forme d'une condition et par plusieurs relations ensemblistes. L'ACI est du type :

**BD1.ELEMENT1**
**<RE> (critère de décomposition)**
**SET( [i, j] BD2.ELEMENT21, [i', j'] BD2.ELEMENT22, …) ∧…∧**

**BD1.ELEMENT1**
**<RE> sinon**

**BD2.ELEMENT23**

Plusieurs critères de décomposition peuvent être enchaînés. Ainsi, pour les critères de décomposition au niveau des noeuds de la BD TOPO et de GEOROUTE (tableau 3) concernant les pattes d'oie et les ronds-points, l'ACI est la suivante :

$$\text{BDT.NOEUD} \equiv_{\text{patte\_oie(G.SET)} \wedge \text{base\_patte\_oie(G.SET)} < 50\ \text{m}} \text{G.SET([3,3] NOEUD, [3,3] TRONÇON)}$$
$$\wedge\ \text{BDT.SET([3,n] NOEUD, [3,n] TRONÇON)} \equiv_{\text{rond\_point(BDT.SET)} \wedge \text{diamètre\_rond\_point(BDT.SET)} < 30\ \text{m}} \text{G.NOEUD}$$
$$\wedge\ \text{BDT.NOEUD} \equiv_{\text{sinon}} \text{G.NOEUD}$$

Cette ACI utilise les fonctions suivantes :

---

[11] Nombre d'arêtes reliées au noeud.

- patte_oie qui teste si l'ensemble de noeuds et de tronçons forment une patte d'oie,
- base_patte_oie qui renvoie la longueur de la base de la patte d'oie,
- rond_point qui teste si l'ensemble de noeuds et de tronçons forme un rond-point,
- diamètre_rond_point qui renvoie le diamètre du rond-point.

Cette extension des ACI aux spécifications des BDG est surtout utile pour le **contrôle de cohérence** (2.1.4). Elle permet, parmi les représentations différentes des mêmes phénomènes, de discerner les différences dues aux spécifications de saisie, des incohérences (erreur de saisie, mise à jour,…).

#### 4.3.3.3 Intégration des classes en conflits de critères de spécification

Les conflits de critères de spécification sont des conflits qui entraînent d'autres conflits. Ils seront donc traités lors de la résolution des conflits qu'ils causent. Les critères de sélection entraînent des conflits de classification, et les critères de décomposition occasionnent des conflits de décomposition (fragmentation). La résolution de ces conflits a déjà été abordée dans les deux sections précédentes.

Ils peuvent aussi provoquer des conflits de données. L'intégration des données ne pose pas de problème particulier dans le cadre d'une **stratégie multi-représentation**. Par contre, pour une **stratégie mono-représentation,** ces conflits rendent l'intégration des géométries des objets particulièrement ardue (4.3.6.2.1).

### *4.3.4 Résolution des conflits de description n-aires*

Il existe un conflit de description n-aire, quand l'information contenue par un ou plusieurs attributs correspond à l'information apportée par plusieurs attributs (3.2.5.3).

#### 4.3.4.1 Déclaration des conflits de description n-aires

Les conflits de description n-aires peuvent être déclarés par des fonctions, des attributs virtuels. De plus, des relations entre les attributs, différentes de l'égalité (correspondance faible et correspondance disjointe), doivent être ajoutées.

##### *4.3.4.1.1 Déclaration de fonctions*

Pour exprimer les différences entre les valeurs des attributs, des fonctions de correspondance ont été proposées [Larson et al. 89]. Cette technique peut être étendue pour les conflits de descriptions n-aires entre les valeurs des attributs. L'AAC (Avec Attributs Correspondants) est alors du type :

$$\mathbf{y = f(x_1, x_2, \ldots, x_n).}$$

Parmi les fonctions, il faut citer les fonctions standard (somme, minimum, maximum, moyenne, cardinalité) couramment utilisées.

Par exemple, pour la correspondance entre les attributs *nb_voies* de la classe TRONÇON de la BD TOPO et *nb_voies_sens_tronçon*, *nb_voies_sens_inverse* de la classe TRONÇON de la BD CARTO une fonction SOMME peut être employée par l'utilisateur.

BDT.TRONÇON.nb_voies
=
SOMME (BDC.TRONÇON.nb_voies_sens_tronçon, BDC.TRONÇON.nb_voies_sens_inverse)

#### *4.3.4.1.2 Déclaration d'attributs virtuels*

Cependant, pour les attributs énumérés en conflits de description n-aires, des fonctions classiques sont difficilement utilisables. Pour résoudre ce problème, des attributs virtuels sont définis pour faciliter la correspondance entre les attributs. Ces attributs virtuels [Souza dos Santos et al. 94] sont des méthodes permettant de mettre en conformité les valeurs des attributs d'un point de vue sémantique. Ils sont construits à partir des attributs initiaux et des **fonctions de transfert** définies par l'utilisateur :

$$\textbf{Att V = f (Att}_1\textbf{, Att}_2\textbf{, …, Att}_n\textbf{)}$$

Ce type de déclaration est une extension des déclarations d'attribut virtuel de [Dupont 95 b].

Pour l'exemple ci-dessous, qui reprend des attributs des classes TRONÇON des classes BD CARTO Version 2 et GEOROUTE,

| | |
|---|---|
| BD CARTO (Version 2) | **Etat_physique** = enum (revêtue, non revêtue, en construction, **chemin exploitation**, **sentier**)<br>**Vocation_liaison** = (autoroute, grande circulation, liaison régionale, liaison locale, bretelle, **piste cyclable**) |
| GEOROUTE | **Etat_physique** = enum (revêtue, non revêtue, en construction)<br>**Vocation_liaison** = (autoroute, artérielle, distribution, desserte, bretelle, passerelle, escalier, voie rapide urbaine, **chemin exploitation ou sentier**) |

Trois attributs virtuels peuvent être créés pour les deux classes :

| | |
|---|---|
| BD CARTO (Version 2) | **Etat_physique_V** = enum (revêtue, non revêtue, en construction)<br>**Type_V** = enum (route, chemin exploitation, sentier, piste cyclable)<br>**Vocation_liaison_V** = (autoroute, grande circulation, liaison régionale, liaison locale, bretelle) |
| GEOROUTE | **Etat_physique_V** = enum (revêtue, non revêtue, en construction)<br>**Type_V** = enum (route, chemin exploitation ou sentier)<br>**Vocation_liaison_V** = (autoroute, artérielle, distribution; desserte, bretelle, passerelle, escalier, voie rapide urbaine) |

les fonctions de transfert sont évidentes, voici quelques exemples significatifs, pour la BD CARTO Version 2 :

$f_{Etat\_physique\_V}$ : Etat_physique × Vocation_liaison → Etat_physique_V
$f_{Etat\_physique\_V}$(« revêtue », « autoroute ») = « revêtue »,
$f_{Etat\_physique\_V}$(« chemin exploitation », « liaison locale ») = «non revêtue»[12],
$f_{Etat\_physique\_V}$(« revêtue », « piste cyclable ») = « revêtue »,

$f_{Type\_V}$ : Etat_physique × Vocation_liaison → Type_V
$f_{Type\_V}$(« revêtue », « autoroute ») = « route »,
$f_{Type\_V}$(« chemin exploitation », « liaison locale ») = « chemin exploitation»,
$f_{Type\_V}$(« revêtue », « piste cyclable ») = « piste cyclable »,

$f_{Vocation\_liaison\_V}$ : Etat_physique × Vocation_liaison → Vocation_liaison_V
$f_{Vocation\_liaison\_V}$(« revêtue », « autoroute ») = « autoroute »,
$f_{Vocation\_liaison\_V}$(« chemin exploitation », « liaison locale ») = « liaison locale »,
$f_{Vocation\_liaison\_V}$(« revêtue », « piste cyclable ») = «»,

Une fois ces attributs virtuels définis, il est possible de déclarer les AAC.

BDC.TRONÇON.Etat_physique_V = G.TRONÇON.Etat_physique_V

#### *4.3.4.1.3 Déclaration de correspondance faible ou disjointe*

Malgré, la déclaration d'attribut virtuel, les attributs peuvent rester incompatibles du point de vue des domaines de définition. Une relation d'égalité ne peut pas être employée entre ces attributs. Pour ces attributs, [Larson et al. 89] ont défini les concepts de correspondance FAIBLE et de correspondance DISJOINTE.

Une **correspondance FAIBLE** est déclarée entre deux attributs : at2 = f (at1) si la valeur de at2 peut être déduite de la valeur de at1 sans qu'il existe d'application bijective.

Une **correspondance DISJOINTE** est déclarée entre deux attributs si la sémantique des attributs est identique sans que la valeur d'un attribut puisse être déduite de la valeur de l'autre. Cependant des **contraintes** portant sur les valeurs peuvent être établies. Elles se révèlent fort utiles pour le contrôle de cohérence.

Pour les attributs *Type_V*, il existe une application injective de la BD CARTO vers GEOROUTE (figure 52), ces deux attributs sont donc en correspondance FAIBLE.

Pour les attributs *Vocation_liaison_V*, les valeurs autorisées sont définies par des organismes différents (IGN, Michelin, DDE) selon la base, ces deux attributs sont donc en correspondance DISJOINTE. Les contraintes définies (figure 53) sont du type « une autoroute de la BD CARTO est soit une autoroute, soit une voie rapide dans GEOROUTE ».

[12] Information déduite de la définition de la valeur chemin d'exploitation

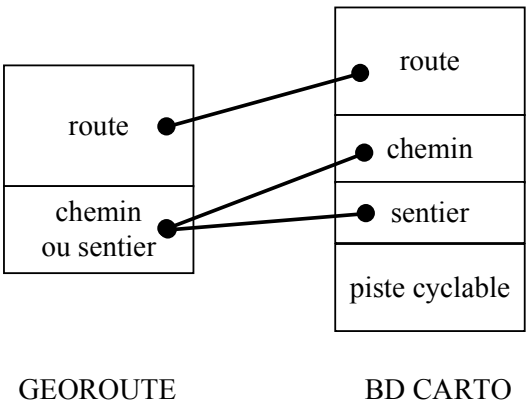



figure 52 : Correspondance FAIBLE entre les attributs Type_V

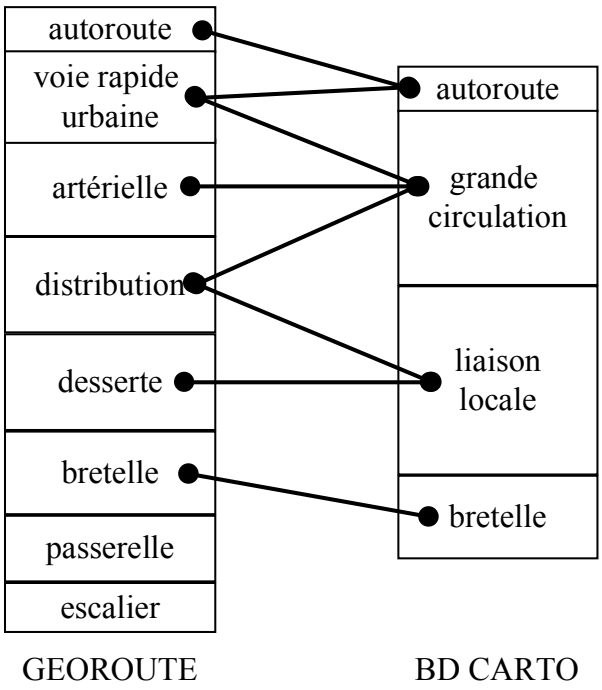



figure 53 : Correspondance DISJOINTE entre les attributs Vocation_liaison_V

Par conséquent, il faut exprimer les correspondances faibles (CF) et disjointes (CD) entre les valeurs possibles des domaines des attributs en correspondance. Ainsi, pour deux attributs énumérés en correspondances at1 et at2 tel qu'at1 = enum (v11, v12,…,v1n) et at2 = enum (v21, v22,…,v2m), les correspondances s'expriment par une AAC :

**at1 CF at2** pour les correspondances faibles

**at1 CD at2** pour les correspondances disjointes

et une **suite de couples**. Chaque couple est composé d'un v1i et d'un ensemble de v2j possibles pour les instances en correspondance. Il n'est pas nécessaire de décrire les correspondances possibles de at2 vers at1 car elles sont induites par les correspondances possibles de at1 vers at2.

Pour la figure 52, on obtient une correspondance faible :

GEOROUTE.TRONÇON.Type_V **CF** BDC.TRONÇON.Type_V
((route,(route)), (chemin ou sentier, (chemin, sentier)))

Pour la figure 53 on obtient une correspondance disjointe :

GEOROUTE.TRONÇON. Vocation_liaison_V **CD** BDC.TRONÇON. Vocation_liaison_V[13]
((autoroute,(autoroute)),
(voie rapide urbaine, (autoroute, grande circulation, liaison local),
(artérielle,(grande circulation)),
(distribution, (grande circulation, liaison locale)),
(desserte,(desserte), (bretelle, (bretelle)),
(passerelle,()),
(escalier, ()))

[13] La valeur « liaison régionale » n'apparaît pas car elle n'a pas été rencontrée dans la base.

#### 4.3.4.2 Intégration des conflits de description n-aires

##### *4.3.4.2.1 Intégration d'AAC avec fonction*

Pour les attributs utilisant une **fonction simple** dans leurs AAC (**att1 = f(att2)**), si cette relation est bijective, alors att1 ou att2 peuvent être indifféremment intégrés. Il n'y a pas de perte d'information, car l'autre attribut peut être déduit en utilisant la fonction f ou $f^{-1}$. Par contre si la fonction n'est pas bijective, l'attribut le plus détaillé est intégré. De même, il n'y a pas de perte d'information, car l'autre attribut (att1) peut être déduit en utilisant la fonction f.

Pour **les fonctions n-aires** ( att1 = f(att21,att22,…)) les attributs intervenant dans la fonction seront ajoutées à la BD intégrée.

##### *4.3.4.2.2 Intégration d'AAC avec attributs virtuels*

De même, l'intégration **d'attributs virtuels** (Att V = f ($Att_1$, $Att_2$, …, $Att_n$)) ne pose pas de problème particulier. A travers la déclaration d'attributs virtuels, des attributs incompatibles sont rendu compatibles. Effectivement, une fois les attributs virtuels déclarés et le mécanisme de transfert défini, les attributs virtuels, s'intègrent comme n'importe quels attributs. Par exemple, l'intégration de BDC.TRONÇON.*Etat_physique_V* et de G.TRONÇON.*Etat_physique_V*, donnera un attribut *Etat_physique* dans le schéma intégré.

##### *4.3.4.2.3 Intégration d'AAC en correspondance faible ou disjointe*

L'intégration d'attributs énumérés en **correspondance faible ou disjointe** est plus compliquée. La technique classique non optimum consiste à garder les deux attributs. Pour rendre l'intégration optimale en terme de redondance, nous proposons une technique de **fusion**. Ces attributs étant des critères potentiels de classification, la technique de fusion des attributs qui va être décrite est semblable à la technique de fusion proposée pour résoudre les conflits de classification.

Pour fusionner les attributs en **correspondance faible,** l'attribut ayant le domaine le plus détaillé est sélectionné dans la BD intégrée. Les valeurs prises par l'autre attribut pouvant être déduites, il n'y aura pas de perte d'information.

Pour l'attribut *Type_V* de la classe TRONÇON de GEOROUTE et de la BD CARTO V2, un attribut *Type* = enum (route, chemin, sentier, piste cyclable) est défini.

Cette solution n'est cependant valable que pour des ACI ayant des relations ensemblistes du type $\supset$, $\supseteq$ ou $\equiv$. Sinon, pour les instances sans correspondant de la BD ayant l'attribut le moins détaillé, la valeur de l'instance correspondante dans la BD intégrée ne peut pas être déterminée. Pour résoudre ce problème, de nouvelles valeurs moins précises doivent être ajoutées au domaine de l'attribut intégré.

Ainsi, pour les instances de la classe TRONÇON de GEOROUTE sans correspondant dans la BD CARTO et prenant la valeur « sentier ou chemin » pour l'attribut Type_V, la valeur de *Type* de l'instance correspondante dans la BD intégrée ne peut pas être définie. La valeur « sentier ou chemin » est donc ajoutée aux valeurs possibles de *Type*.

Pour fusionner les attributs en **correspondance disjointe**, un attribut ayant pour valeurs possibles le **produit cartésien** des valeurs des domaines initiaux est défini. Pour des attributs att1 et att2 ayant respectivement les domaines (v11,v12,…) et (v21,v22,…), les valeurs possibles du domaine résultant sont les v1i-v2j. Le produit cartésien est une solution

envisageable pour une ACI ayant pour relation ensembliste l'équivalence. Pour les autres ACI, il est nécessaire de rajouter parmi les valeurs possibles, les valeurs v1i- ou v2i-.

Cependant, le produit cartésien n'est pas la solution optimale pour des correspondances 1-n alternées. Une correspondance **1-n alternée** existe si l'on peut définir deux sous-domaines tels que, pour chaque couple de sous-domaine en relation, il existe toujours soit une relation 1-n, soit une relation n-1 au niveau des valeurs. Dans ce cas, on peut intégrer ces attributs en un attribut prenant comme domaine l'union des domaines les plus détaillés.

| |
|---|
| att1 = enum (A,BC,D,E,F)<br>att2 = enum (A,B,C,DEF)<br>att1 **CD** att2<br>(A, (A)), (BC,(B,C)), (D, (DEF)), (E, (DEF)), (F, (DEF)),<br>**⇒ attI = enum (A,B,C,D,E,F)** |

tableau 7 : Exemple d'attributs en correspondance 1-n alternée

Pour l'exemple du tableau 7, l'attribut att1 est plus précis au niveau des valeurs D,E et F, et att2 est plus précis au niveau des valeurs B et C. Les attributs att1 et att2 sont en relation disjointe. Cependant des sous-domaines (A, BC) et (D, E, F) pour att1 et (A, B, C) et (DEF) pour att2, peuvent être définis tel que (A, BC) de att1 et (A, B, C) de att2 soient en relation n-1 et (D, E, F) de att1 et (DEF) de att2 soient en relation 1-n. Nous sommes donc en présence d'une relation 1-n alternée. Ces deux attributs peuvent être fusionnés en un attribut énuméré att1 prenant comme valeurs possibles (A, B, C, D, E, F). Cette solution est valable pour une ACI ayant une relation ensembliste d'équivalence. Pour les autres types de relation ensembliste, les problèmes décrits pour le produit cartésien se posent à nouveau et les extensions proposées sont similaires.

### *4.3.5 Résolution des conflits de granularité*

Il existe un conflit de granularité entre deux classes, si deux attributs en correspondance ont des granularités différentes (3.2.3.3.2).

#### 4.3.5.1 Déclaration des contraintes de granularité

Les AAC doivent pouvoir inclure la notion de granularité pour distinguer les conflits de données dus à une incohérence (erreur, mise à jour non effectuée) des conflits engendrés par des granularités différentes. La différence de granularité d'un attribut est exprimée dans l'AAC par un prédicat renvoyant « vrai » si une mesure sur un élément (longueur pour un objet linéaire, surface pour un élément surfacique) est supérieure à un seuil. Ce prédicat est du type :

$$\mathbf{BD1.el1.att1 =}_{\mathbf{L(el1) > x\ unité}}\ \mathbf{BD2.el2.att2} \qquad \mathbf{el1\ linéaire}$$

$$\mathbf{BD1.el1.att1 =}_{\mathbf{S(el1) > x\ unité}}\ \mathbf{BD2.el2.att2} \qquad \mathbf{el1\ surfacique}$$

Pour l'exemple de la figure 30 (page 3), le nombre total de voies d'un tronçon de GEOROUTE est égal au nombre total de voies du tronçon équivalent de la BD CARTO si la longueur du tronçon de GEOROUTE est supérieure à 1000 m. L'AAC suivante est donc obtenue :

$$\text{G.TRONÇON.nb\_voies} =_{\text{L(G.TRONÇON) > 1000 m}} \text{BDC.TRONÇON.nb\_voies}$$

#### 4.3.5.2 Intégration des conflits de granularité

Pour une stratégie **multi-représentation,** les conflits de granularité entraînent une relation de composition entre un objet de la classe ayant une granularité grossière pour cet attribut et un ou plusieurs objets de la classe ayant une granularité fine pour le même attribut.

Pour intégrer l'attribut en conflit, un attribut correspondant est créé dans la classe composante. La valeur de cet attribut pour la classe composée est renvoyée par une **méthode** qui la calcule en fonction des valeurs prises par cet attribut pour les objets composants.

T_généralisé ⇔(t1,t2,t3) avec
t1.nb_voies = 2, t2.nb_voies = 4, t3.nb_voies = 2
longueur(t1) = 3 234 m, longueur(t2) = 746 m, longueur(t3) = 1 484 m
granularité de la classe de T_généralisé pour l'attribut nb_voies = 1 000 m
alors la méthode nb_voies de la classe de T_généralisé renvoie
T_généralisé.nb_voies = 2

Dans l'exemple ci-dessus, dans la BD intégrée, le tronçon T_généralisé correspond aux tronçons t1, t2 et t3. La classe de T_généralisé n'a pas d'attribut nb_voies mais une méthode renvoyant le nombre de voies. Cette méthode recherche les valeurs de l'attribut nb_voies (2,4,2) des objets composants (t1, t2, t3). En cas de différence entre les valeurs, seules les valeurs des objets ayant une longueur supérieure à la granularité (1 000 m) seront sélectionnées (2,2). Les valeurs seront alors obligatoirement similaires.

Cependant, cette technique d'intégration des attributs en conflit de granularité doit être limitée aux classes dont l'ensemble des attributs en conflit de granularité est inclus dans la même classe.

Pour les autres cas, des correspondances n-m sont engendrées au niveau des instances. Les valeurs des attributs des classes composées ne peuvent plus être déduites. Il faut alors conserver les attributs en conflit de granularité dans les deux classes.

Pour une stratégie **mono-représentation**, les conflits de granularité posent exactement les mêmes problèmes que les autres conflits de fragmentation et entraînent des scissions des géométrie des objets (4.3.6.2.1.1).

### *4.3.6 Résolution des conflits de description de la géométrie pour les données vecteurs*

#### 4.3.6.1 Déclaration de conflits de description entre les attributs géométriques

Pour déclarer les correspondances entre les géométries une dernière clause :**« Avec Géométrie Correspondante » (AGC)** est ajoutée. Elle est employée chaque fois qu'il est possible de spécifier comment les géométries des BDG sont en relation.

Par exemple, pour la figure 31 (page 3), si la BD1 est obtenue en généralisant la BD2, la clause AGC est la suivante :

AGC : BD1.TRONÇON.géométrie = MERGE ( UNSPLIT ( {x.géométrie / x ∈SET(BD2.TRONÇON, BD2.SEPARATEUR)}

La valeur de la géométrie d'une instance de la classe TRONÇON de la BD1 peut être calculée à partir de la géométrie des instances correspondantes des classes TRONÇON et SEPARATEUR, grâce aux fonctions UNSPLIT qui groupent les lignes adjacentes en une

ligne continue et MERGE qui fusionne ces lignes continues quasi parallèles en une ligne unique médiane.

Néanmoins, les géométries des bases à intégrer sont rarement dérivées les unes des autres. Elles sont issues de saisies différentes et répondent à des spécifications différentes. Il est donc nécessaire d'affaiblir l'égalité entre les valeurs des géométries en ajoutant une **tolérance**, qui est fonction de l'exactitude et de la résolution.

Deux types de tolérance sont possibles :
- la tolérance maximale, qui va être employée pour vérifier que la distance maximale entre les deux géométries est inférieure à cette tolérance.
- la tolérance moyenne, qui va être employée pour vérifier que la distance moyenne entre les deux géométries est inférieure à cette tolérance.

Ainsi, pour ce même exemple, si les géométries des BDG sont issues de saisies différentes l'AGC avec une tolérance moyenne $D_m$ peut être la suivante :

AGC : $D_m$ (BD1.TRONÇON.géométrie, MERGE ( UNSPLIT ( {x.géométrie / x ∈ SET(BD2.TRONÇON, BD2.SEPARATEUR)}) < BD1.exactitude + BD2.exactitude

Les AGC avec tolérance, sont surtout nécessaires dans le contexte du contrôle de cohérence.

#### 4.3.6.2 Intégration des géométries en conflit

L'intégration des géométries est une tâche complexe. Les deux stratégies s'opposent radicalement pour résoudre ce problème.

##### *4.3.6.2.1 Stratégie mono-représentation*

Pour la stratégie mono-représentation**,** les géométries à unifier sont **fusionnées** en une seule géométrie. Si les deux BDG à intégrer représentent des phénomènes distincts et des couches géométriques de type Spaghetti, cette fusion correspond à une simple superposition (overlay).

Par contre, si les deux bases représentent des phénomènes similaires, il faut définir une unique géométrie pour chaque objet représentant le même phénomène. Mise à part les techniques de superposition [Frank 87] [Dougenik 80] [Pullar 93], très peu de techniques de fusion [Ubeda et Egenhofer 97] ont été développées pour résoudre les différences entre les deux géométries. Pour contourner ce problème, nous avons choisi de constituer la géométrie de la BDI 1 à partir de la BD TOPO qui a été modifiée lors de la pré-intégration. La géométrie de GEOROUTE est retenue uniquement pour les différences permettant d'enrichir la BDI 1 et doit être intégrée en :
- scindant les géométries de la BD TOPO pour tenir compte de la richesse sémantique de la GEOROUTE,
- migrant les pattes d'oie de petite base de GEOROUTE
- migrant les impasses de GEOROUTE sans correspondant.

Ces trois enrichissements vont être présentés maintenant. Ils sont propres aux BD de l'IGN mais peuvent servir d'exemple pour d'autres BDG.

###### *4.3.6.2.1.1 Scission*

Les tronçons de GEOROUTE ont des attributs propres. Entre deux carrefours, la valeur prise par ces attributs peut changer : il existe alors un noeud routier de type « changement d'attribut » et deux tronçons. Par contre, pour la BD TOPO qui ne possède pas cet attribut, un seul tronçon est présent sur la même zone. Or, les attributs propres à GEOROUTE sont

présents dans la base intégrée BDI 1. Il faut donc scinder le tronçon issu de la BD TOPO dans la BDI 1 et créer un noeud routier correspondant au noeud routier de GEOROUTE reliant ses deux tronçons [Ousset 97]. Pour positionner ce noeud de la BDI 1, le **rapport des longueurs** est employé :

$$L(BDI1.Ti) = L(BDT.T)\frac{L(G.Ti)}{\sum_i L(GTi)}$$

Il permet de définir la longueur de chaque tronçon issu de la scission.

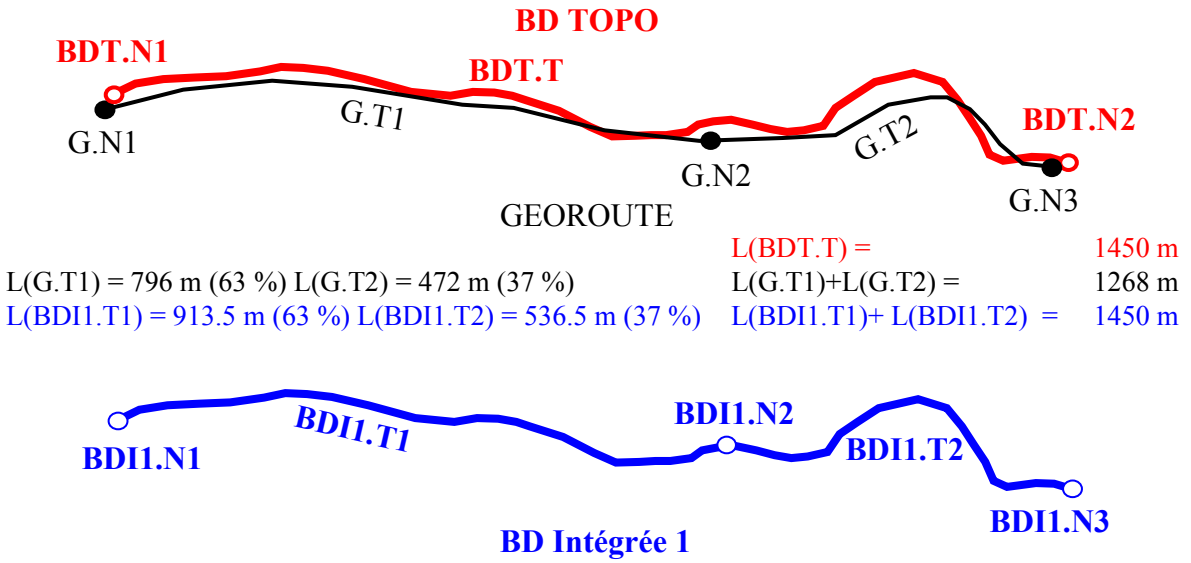



figure 54 : Scission d'un tronçon BD TOPO due à un changement de valeur d'un attribut propre à GEOROUTE

Ainsi, pour l'exemple de la figure 54, la longueur de G.T1 représente 63 % de la longueur des tronçons G.T1 et G.T2 qui correspondent à BDT.T. Dans la BD intégré 1, BDT.T1 est scindé en deux tronçons (BDI 1.T1, BDI 1.T2) tels que la longueur de BDI 1.T1, le tronçon qui correspond à G.T1, soit égale à 63% de la longueur de BDT.T.

D'autres techniques peuvent être réalisées pour calculer le point de scission, comme la recherche du **point le plus proche** sur le tronçon BD TOPO du noeud GEOROUTE. Cette technique permet une localisation absolue la plus précise, mais n'est pas satisfaisante du point de vue de la localisation relative et du rapport des longueurs (figure 55). Elle n'a donc pas été retenue.


figure 55 : Comparaison du point le plus proche et
du point conservant le rapport des longueurs

Une fois les tronçons et les noeuds définis, il faut affecter des valeurs à leurs attributs et à leurs relations. Cette opération est similaire à celle définie lors de la pré-intégration pour la transformation des tronçons de chaussée en tronçon de route (4.2.3).

#### *4.3.6.2.1.2 Migration des pattes d'oie*

Les pattes d'oie de 20 à 50 mètres de base (l'écartement) sont saisies par deux tronçons dans GEOROUTE et par un tronçon dans la BD TOPO. Pour obtenir l'information la plus précise, dans la BD intégrée, il faut migrer ces pattes d'oies dans la BDI 1. Etant donnée leur courte distance, les tronçons des pattes d'oie de 20 à 50 m sont représentés dans la BDI 1 par un seul segment de droite. Pour ajouter les pattes d'oie dans la BDI 1, un triangle similaire ayant la même base (b) et la même hauteur (h) remplace la fin du tronçon issue de la BD TOPO (figure 56).

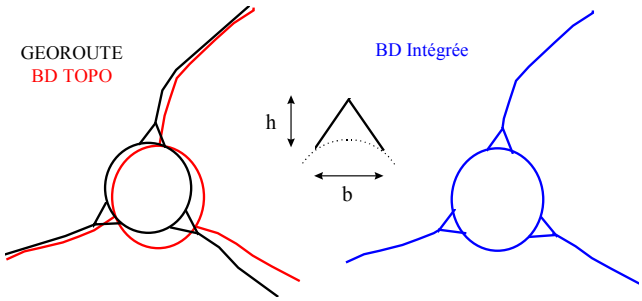



figure 56 : Ajout des pattes d'oies de GEOROUTE dans la BD intégrée

#### *4.3.6.2.1.3 Intégration des impasses*

L'intégration des impasses de GEOROUTE sans correspondant n'est pas triviale. Premièrement, il faut scinder le tronçon correspondant au tronçon où débouche l'impasse, la technique employée est la même que pour la scission. Deuxièmement, il faut reporter l'impasse en elle-même. La solution naïve consiste à reporter le tronçon à partir de sa position relative vis-à-vis du nouveau noeud (figure 57). Cette solution risque de poser des problèmes dus à la position relative des objets, l'insertion d'une impasse risque de créer des superpositions au niveau du réseau (tronçon qui s'intersectent) ou vis-à-vis d'un objet d'une autre classe (tronçon passant sur une habitation). Il n'y a pas de solution automatique à ces problèmes, il faut modifier manuellement la géométrie de l'impasse de telle sorte qu'il n'y ait plus de conflits.

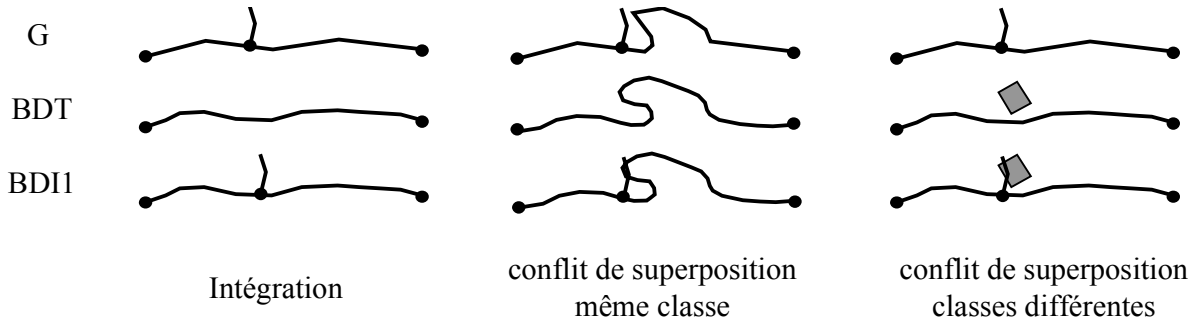



figure 57 : Conflits entraînés par l'intégration des impasses de GEOROUTE dans la BD intégrée 1

#### *4.3.6.2.2 Stratégie multi-représentation*

Pour intégrer les géométries, la technique utilisée pour la stratégie multi-représentation est la **préservation**. Les deux couches géométriques sont donc conservées. Par contre, les objets géographiques s'y rapportant sont intégrés et bénéficient de deux géométries (la géométrie de la première couche et la géométrie de la deuxième couche).

Ainsi, pour la figure 58, le vieux bois a une géométrie dans chaque couche de la BD intégrée. Les relations entre les objets géographiques et les géométries peuvent être plus complexes du fait des conflits de fragmentation. Un objet généralisé peut être décomposé en plusieurs objets. La géométrie de l'objet généralisé correspond alors à plusieurs géométries. D'une manière plus générale, plusieurs géométries correspondent à plusieurs géométries. Les liens entre les géométries sont alors indirectement établis via les objets.

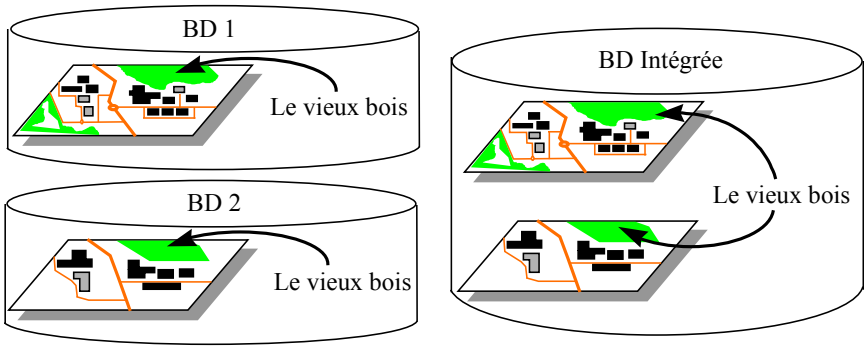



figure 58 : Exemple de préservation de la géométrie

## 4.4 Conclusion

### *4.4.1 Contribution du processus d'intégration de BDG*

Ce chapitre a montré qu'il était possible d'étendre les processus d'intégration classiques aux BDG. Pour cela, nous avons repris chacune des 3 phases et nous les avons enrichies.

La **pré-intégration** a été élargie en ajoutant :

- des mécanismes de conversion,
- des méta-données nécessaires à la manipulation de données hétérogènes,
- des données implicites qui ont été matérialisées,
- trois règles de normalisation spécifiques aux BDG.

La syntaxe des **déclarations de correspondance** a été étendue afin de permettre la déclaration des correspondances entre les BDG et de leurs conflits spécifiques (conflit de spécification, conflit de granularité, …). Elle permet l'ajout :

- de la notion de direction nécessaire à la déclaration des correspondances entre attribut dépendant de la direction,
- de la clause « appariement géométrique des données » (AGD), pour identifier les objets homologues,
- d'ensemble d'éléments (classe, relation) pour déclarer les conflits de classification,
- de classe virtuelle nécessaire pour déclarer les conflits de fragmentation,
- de sélection exprimant les critères de sélection,
- de relation ensembliste conditionnée par un critère de décomposition,
- d'attribut virtuel permettant de mettre en conformité sémantique les attributs,
- de correspondance faible entre attributs pour les correspondances entre des attributs ayant uniquement une fonction injective les reliant,
- de correspondance disjointe entre attributs pour déclarer des correspondances entre des attributs ayant la même sémantique mais aucune fonction injective ou surjective les reliant,
- de fonctions géométriques (MERGE, SPLIT,…) permettant de relier les attributs géométriques,
- de tolérance entre les valeurs des géométries.

Ces extensions sont novatrices et proposent une syntaxe relativement simple et intuitive.

Pour intégrer les BDG, nous avons commencé par définir deux **stratégies** (stratégie mono-représentation et stratégie multi-représentation) qui ont des objectifs très différents. Le résultat de la première stratégie est une représentation unique des phénomènes du monde réel. Tandis que la deuxième conserve les représentations à différents niveaux de détail et les relie entre elles. Pour chaque conflit et chaque stratégie, des techniques d'intégration ont pu être définies :

- Pour les conflits de **classification**, les techniques classiques de fusion et de partition ont été retenues. Toutefois, la technique de fusion a été enrichie par l'ajout d'un attribut énuméré de classification, ce qui permet de conserver l'information initiale.
- Pour résoudre les conflits de **fragmentation,** une technique innovante s'appuyant sur des relations de composition et des classes virtuelles a été avancée.

Pour intégrer les **attributs**, de nouveaux mécanismes ont aussi été développés. Ils tiennent compte des spécificités des BDG dont la direction. Pour intégrer des attributs de **granularité** différentes, des méthodes ont été utilisées pour représenter les attributs ayant la granularité la plus grossière. De même, pour les **attributs énumérés** en correspondance disjointe, la recherche de correspondance 1-n alternée a permis d'optimiser la fusion de ces attributs.

Pour intégrer les **géométries**, une fusion des géométries a été réalisée, dans le cadre d'une stratégie mono-représentation. Elle s'appuie sur la géométrie la plus détaillée et intègre uniquement les géométries de l'autre BDG si elle permet d'enrichir la représentation intégrée. Dans cette objectif, des techniques de scission et de migration ont été définies. Par contre, pour la stratégie multi-représentation, les différentes géométries ont simplement été reliées via les objets géographiques.

Les **conflits** pour les BDG vecteur, ont ainsi pu être résolus.

Ce processus d'intégration a été défini en collaboration avec le département d'informatique de l'EPFL [Devogele et al. 97] et il a été validé pour les BDG de l'IGN en utilisant $GéO_2$.

L'ensemble des ACI des BD de l'IGN est donné en Annexe 7.7 ainsi que les schémas intégrés de la BDI 1 et de la BDI 2 en annexe 7.7.1.

### *4.4.2 Perspectives pour le processus d'intégration de BDG*

Le processus d'intégration que nous avons défini, permet d'intégrer le thème routier des BDG de l'IGN. Il peut cependant être amélioré en :

- assistant la déclaration des correspondances,
- améliorant la technique d'intégration des géométries de la stratégie mono-représentation,
- améliorant la technique d'intégration des attribut en conflits n-m de la stratégie mono-représentation,
- approfondissant l'intégration des relations,
- en rendant automatique la phase d'intégration.

La déclaration des correspondances peut être rendue moins ardue **en assistant l'utilisateur dans sa recherche des correspondances**. Pour les BD classiques, des prototypes permettant **d'analyser les similarités** (nom, domaines de valeurs, cardinalité, etc.), entre les éléments des schémas ont déjà été développés [Hayne et Ram 90] [Gotthard et al. 92], des outils analogues peuvent être développés pour les BDG. Une autre piste consisterait à rechercher les correspondances en fonction des méthodes des classes [Kuhn 94], ceci afin de réduire les problèmes liés à l'attribution de nom pour les classes et les attributs (homonyme, synonyme). Cependant, dans tous les cas, la décision finale, pour décider si cette correspondance induite est fondée appartient à l'intégrateur en fonction de la sémantique des données.

Pour la stratégie mono-représentation, **l'intégration des géométries** n'est pas parfaite. Entre autres, la solution proposée pour intégrer des pattes d'oies n'est pas optimale. Elle pourrait être améliorée en utilisant des mécanismes permettant de conserver la forme de la patte d'oie de Géoroute dans la BD intégrée (par exemple la transformation élastique (Rubber-sheeting) et la triangulation [Fagan et Soehngen 87][Chikh-Zaghar 94] [Shmutter et Doytsher 92] [Gillman 85]). L'intégration de géométrie demande aussi la définition de contraintes de localisation (superposition avec un autre objet possible ou non, position relative autorisée, …) [Ubeda et Egenhofer 97] permettant de juger si l'intégration proposée est acceptable ou si une autre géométrie doit être déterminée.

De plus, si la qualité géométrique des deux bases est similaire, une meilleure géométrie pourrait être obtenue, en utilisant les géométries des deux bases et non à partir de la géométrie d'une base posée en référence. La détermination des géométries résultantes est alors complexe, qui nécessite d'améliorer les processus de superposition (Overlay) existants. Une tentative de fusion est décrite dans [Ubeda et Egenhofer 97].

Pour cette stratégie mono-représentation, des améliorations sont aussi nécessaires pour **intégrer les attributs** en présence d'un conflit n-m au niveau des objets. Elles devront régler tous les problèmes liés à la désagrégation de valeurs aussi appelée interpolation [Flowerdew et Openshaw 87] [Weber 94] sans dégrader les données.

L'**intégration des relations** n'a été qu'évoquée dans cette thèse et mériterait une étude plus approfondie. Effectivement, cette intégration est fortement contrainte par l'intégration des classes (deux liens peuvent être intégrés uniquement si les objets liés sont déjà intégrés). De plus, l'intégration des relations doit aussi tenir compte des conflits au niveau des classes reliées par ces relations. Ainsi les liens d'une relation reliant des instances regroupées dans une instance virtuelle (conflits de fragmentation), n'ont pas de lien correspondant dans l'autre base. Par contre, les conflits au niveau des relations sont proches des conflits au niveau des

classes. Une description de ces conflits, de leur déclaration et de leur résolution est disponible dans [Dupont 95 b]. Pour les BDG, l'intégration des relations en correspondance sans conflit, dépend de la stratégie. Pour une stratégie **mono-représentation**, les instances des classes ont déjà été fusionnées. Les liens entre ces instances peuvent être aussi fusionnés. Par contre, pour une stratégie multi-représentation, la technique d'intégration dépend de l'intégration des classes réalisées. Si les classes reliées ne sont pas en conflit de fragmentation, les relations peuvent être fusionnées. Inversement, si les classes sont en conflit de fragmentation, des relations de compositions sont créées, la fusion des relations est alors impossible. Deux solutions sont envisageables, la plus simple consiste à préserver les relations, cette technique n'est pas à proprement parler une technique d'intégration. La deuxième consiste à déduire les relations (relation virtuelle définie par des méthodes) de la représentation la moins détaillée à partir des relations de la représentation la plus détaillée.

Actuellement, la phase d'intégration est manuelle. Or, pour chaque conflit et stratégie, une technique d'intégration a été définie pour intégrer les classes, l'intégration des classes peut donc être largement **automatisée**.

Plus généralement, l'intégration des BDG doit aussi être élargie :

- aux BDG ayant des modèles différents. Pour résoudre ce conflit, des mécanismes de traduction entre modèles doivent être déterminés,
- aux BDG raster et vecteur,
- aux BDG incluant la troisième dimension (l'altitude),
- aux méthodes et fonctions des SIG des BDG,
- aux relations d'héritage des classes des BDG,
- aux BDG incluant la dimension temporelle,

Pour conclure, ce processus d'intégration a montré que l'enrichissement des déclarations de l'identification des éléments communs est complexe et ne peut pas toujours être déclaré de manière individuelle. Les déclarations individuelles doivent alors être remplacées globalement par un processus d'appariement. Ce processus d'appariement est abordé dans la partie suivante.

# 5. Appariement de BDG

Pour rappel, **l'appariement** encore appelé **conflation** est le processus consistant à établir les correspondances entre les objets géographiques des différentes bases qui représentent le même phénomène du monde réel. Il est utilisé dans de nombreuses applications manipulant l'information géographique : regroupement de bases de données juxtaposées [Laurini 96], propagation des mises à jour dans une base de données client [GIS/Trans Ltd 95] [Bucaille 97], recalage de données sur un référentiel [Lupien et Moreland 87] [Lynch et Salford 85], intégration de BDG [Gouvernement du Québec 92], contrôle qualité [Brooker 95], superposition de couches pour fusionner les géométries [Schorter et al. 94].

Pour intégrer des BDG, une clause **appariement géométrique des données (AGD)** qui spécifie le prédicat d'appariement entre les instances peut être définie. Cependant, elle peut rarement être employée du fait de la difficulté à identifier les objets homologues. Une alternative consiste à remplacer l'ensemble des clauses AGD par un processus d'appariement global. Ce processus est exécuté parallèlement à l'intégration des schémas des BDG.

Dans l'état de l'art, plusieurs outils pouvant être utilisés pour apparier des objets géographiques, ont été présentés. Ils s'appuient sur :

- la comparaison de la sémantique des objets,
- l'utilisation de distances entre leur géométrie,
- la ressemblance de leur forme,
- les correspondances entre leurs relations topologiques.

Cependant, aucun outil pris individuellement n'est suffisant. Effectivement, les objets géographiques des différentes bases représentant le même phénomène du monde réel, présentent des différences importantes (conflits de données). Afin d'obtenir un résultat fiable, il faut donc combiner un grand nombre d'outils et s'appuyer sur les résultats d'appariements déjà réalisés pour en déterminer de nouveaux.

Les outils d'appariement sélectionnés vont renvoyer, en plus des objets à apparier, des objets dit **parasites** (objets répondant aux critères définis par l'outil d'appariement mais ne devant pas être appariés). Il faudra donc compléter les outils d'appariement par des outils de **filtrage** permettant de restreindre le résultat aux éléments devant être appariés.

L'appariement est un processus complexe. Pour faciliter la conception des processus d'appariement, un processus générique d'appariement s'appuyant sur une boîte à outils, sera présenté dans le chapitre 5.1. Puis, afin de l'illustrer, le prototype d'appariement réalisé pour les données routières de GEOROUTE et de la BD CARTO sera exposé et les résultats analysés dans le chapitre 5.2. Enfin, des enrichissements possibles pour la BD intégrée, à l'aide du processus d'appariement, seront avancés dans le chapitre 5.3.

## 5.1 Processus générique s'appuyant sur une boîte à outils d'appariement

Les techniques d'appariement n'en sont encore qu'à leur début. Les besoins sont présents et nombreux, mais les outils sont pour l'instant insuffisants et chaque application répond à des problèmes spécifiques.

Cependant, à partir de l'état de l'art et des processus d'appariement développés au laboratoire COGIT, un squelette de processus générique s'appuyant sur une boîte à outils d'appariement peut être défini. Quel que soit le type de données (réseau, occupation du sol,…) et quelles que

soient les différences entre les données, ce squelette servira de structure pour concevoir les processus d'appariement spécifiques.

Le **but** de ce chapitre est donc de décrire un module générique d'appariement qui regroupe les étapes communes aux processus d'appariement. Ce squelette permettra de construire des processus d'appariement complexes sans développer à nouveau l'ensemble du processus générique. Ainsi, l'utilisateur pourra se concentrer uniquement sur les choix des outils d'appariement pour apparier ses données. Les processus d'appariement gagneront de la sorte en qualité et en performance.

Le processus générique d'appariement s'appuie sur une boîte à outils qui sera présentée en 5.1.1. Il est composé de 6 étapes :

- l'enrichissement (5.1.2.1),
- la sélection des objets candidats à l'appariement (5.1.2.2),
- le calcul des mesures d'appariement (5.1.2.3),
- le filtrage ou la prolongation (5.1.2.4),
- le regroupement d'appariement (5.1.2.5),
- l'analyse du résultat (5.1.2.6).

L'enchaînement des étapes est décrit en 5.1.2.7.

## *5.1.1 Définition d'une boîte à outils*

Un **outil d'appariement** est une application intervenant dans un processus d'appariement. Afin de faciliter leur choix, les outils d'appariement doivent être regroupés en fonction de l'information, caractéristique qu'ils mesurent, et être renseignés par des contraintes d'utilisation et le type de liens qu'ils renvoient.

Nous avons défini cinq groupes d'outils (outils sémantiques, outils géométriques de distance, outils géométriques de forme, outils topologiques, outils mixtes). L'ensemble des outils proposés est présenté en annexe 7.3. Seuls les outils qui n'ont pas été cités dans l'état de l'art et les nouveaux outils développés pour notre prototype routier sont présentés dans cette section.

### 5.1.1.1 L'outil sémantique

Pour apparier des données géographiques selon leur sémantique, un outil d'appariement sémantique doit être présent. L'**égalité sémantique** a été retenue.

| |
|---|
| **egalite_semantique** ( objets1 : set(objet_bd1), propriétés1 : list(string),<br>objets2 : set (objet_bd2), propriétés2 : list(string)) :<br>set (n-uplet ( appariés1 : set(objet_bd1),<br>appariés2 : set(objet_bd2))) |

L'égalité sémantique est l'opération qui teste l'égalité entre deux ensembles d'objets au niveau des valeurs renvoyées par un ensemble de propriétés (valeur d'un attribut ou valeur renvoyée par une méthode). Elle retourne une suite de couples composés d'un ensemble d'objets de la première base et d'un ensemble d'objets de la deuxième base, appariés sémantiquement.

| Objet 1 | propriété $_{11}$ | propriété $_{12}$ |
|---|---|---|
| a | 1 | 2 |
| b | 2 | 2 |
| c | 3 | 1 |
| d | 4 | 1 |
| e | 4 | 1 |

| Objet 2 | propriété $_{21}$ | propriété$_{22}$ |
|---|---|---|
| A | 1 | 2 |
| B | 2 | 2 |
| C | 2 | 2 |
| D | 4 | 1 |

Pour l'exemple ci-dessus,

egalite_semantique ((a, b, c, d, e), (A, B, C, D), (propriété$_{11}$, propriété$_{12}$), (propriété$_{21}$, propriété$_{22}$))

l'égalité sémantique va fournir le résultat suivant

( ( (a), (A)), ( (b), (B,C)), ( (d, e), (D))

Cet outil a l'avantage d'être souple et générique. En effet, les objets des ensembles passés en paramètre, peuvent être des instances de classes différentes (par exemple TRONÇON_ROUTE, TRONÇON_SENTIER), ainsi ils ne subissent pas les conflits de classification (3.2.3.1). De même, cette fonction peut prendre en paramètres des méthodes et des attributs afin de pouvoir définir des **attributs virtuels** (4.3.4.1.2) dont la valeur est calculée à partir des valeurs des attributs des classes.

Enfin, il est indépendant des classes des objets et du type des valeurs renvoyées (entier, réel, énuméré, chaîne de caractères, …). Les liens générés par cet outil peuvent être aussi bien des liens 1-0 que des liens n-m (sa seule contrainte est l'égalité des domaines de valeur entre la *propriétés1[i]* et la *propriétés2[i]*). Un seul outil d'appariement sémantique est donc suffisant.

#### 5.1.1.2 Les outils géométriques de type distance

En revanche, plusieurs outils géométriques doivent être définis du fait de l'imprécision de la géométrie et des nombreux conflits qui y sont liés. Dans l'état de l'art (2.3.2), des distances ont été présentées pour les objets ponctuels (la **distance euclidienne**), pour les **objets linéaires** (la **distance moyenne,** la **distance de Fréchet**, et la **distance de Hausdorff**) et pour les objets surfaciques (la **distance surfacique**).

Ces cinq distances renvoient des mesures de « proximité », elles n'établissent pas d'appariement entre les instances. Elles sont donc utilisables quelle que soit la cardinalité des appariements recherchés.

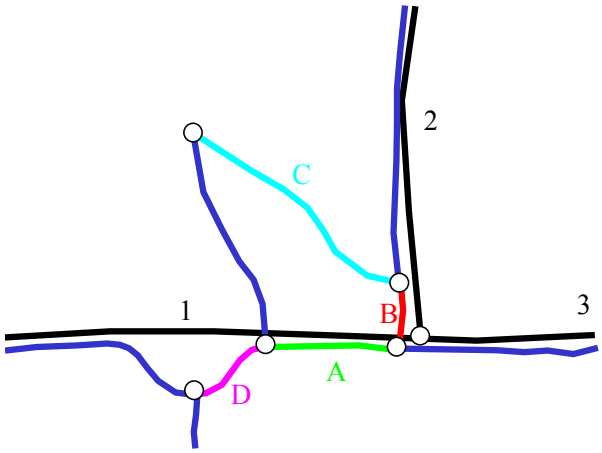
Pour des appariements 1-n, la notion d'objet le plus proche n'est pas pertinente, et les distances maximum (Hausdorff, Fréchet) ne sont pas des outils adaptés car, le plus souvent, la distance sera égale à la distance entre une des extrémités du plus grand des deux tronçons et le point le plus proche de l'autre tronçon (figure 18).

Nous avons donc repris et amélioré la technique d'appariement défini par Stricher [Stricher 93], [Raynal et Stricher 94]. La technique de Stricher reçoit en entrée un ensemble de lignes à comparer, un ensemble de lignes références et un seuil. Elle compare la mesure de la **composante de la distance de Hausdorff** des lignes de la base à comparer (ligne à

comparer) vers les lignes de la base référence avec le seuil[14]. Ainsi, une ligne à comparer sera dite :

- **appariée géométriquement**, s'il existe une seule ligne référence dont la valeur de la composante de la distance de Hausdorff est inférieure au seuil,
- **litigieuse,** s'il existe plusieurs lignes références dont la mesure de la composante est inférieure au seuil,
- **non appariée géométriquement,** s'il n'existe pas de ligne référence dont la mesure de la composante est inférieure au seuil.

Le choix du seuil est donc primordial pour l'appariement géométrique des lignes.




$D_{A\to1}$ = 0,93 m
$D_{A\to2}$ = 9,48 m
$D_{A\to3}$ = 9,48 m
$D_{B\to1}$ = 3,45 m
$D_{B\to2}$ = 1,04 m
$D_{B\to3}$ = 3,60 m
$D_{C\to1}$ = 12,00 m
$D_{C\to2}$ = 12,90 m
$D_{C\to3}$ = 18,87 m
$D_{D\to1}$ = 4,20 m
$D_{D\to2}$ = 14,13 m
$D_{D\to3}$ = 14,13 m

figure 59 : Mesure de la composante de Hausdorff

| **seuil** | | | | |
|---|---|---|---|---|
| **15 m** | A, (1,2,3) | B, (1,2,3) | C, (1,2) | D, (1,2,3) |
| **12 m** | A, (1,2,3) | B, (1,2,3) | C, () | D, (1) |
| **9 m** | A, (1) | B, (1,2,3) | C, () | D, (1) |
| **6 m** | A, (1) | B, (1,2,3) | C, () | D, (1) |
| **3 m** | A, (1) | B, (2) | C, () | D, () |

tableau 8 : Résultats renvoyés en fonction du seuil pour la figure 59

Un seuil trop grand entraîne la sélection d'un grand nombre d'appariements litigieux (seuil de 9 m et plus pour le tableau 8), un seuil trop bas provoque des non sélections dommageables (seuil de 3 m pour le tableau 8).

Pour sortir de cette impasse, nous avons donc amélioré la technique de Stricher, en procédant par **seuillages successifs** de plus en plus fins pour éliminer au fur et à mesure les lignes litigieuses**.**

Pour l'exemple de la figure 59, un seuil de 9 m puis de 3 m (tableau 9) permet de retenir l'ensemble des appariements à 9 m et de résoudre les appariements litigieux par un deuxième appariement à 3 m. Ainsi, la ligne B qui est litigieuse à 9 m est appariée avec 2 à 3 m. L'appariement géométrique à partir du calcul de la composante de Hausdorff avec deux seuillages donnera donc un résultat plus robuste.

[14] Cette composante de la distance de Hausdorff a été mis en avant par [Abbas 94] comme mettant en évidence l'imprécision ou la généralisation entre les données à comparer. C'est un outil tout désigné lorsque la précision des BDG est différentes.

| seuil | | | | |
|---|---|---|---|---|
| **9 m puis 3 m** | A, (1) | B, (2) | C, () | D, (1) |

tableau 9 : Résultats renvoyés avec un seuillage successif pour la figure 59

La signature de la technique d'appariement de Stricher améliorée est donc :

**Appariement_composante_Hausdorff** (objets_à_comparer : set(objet_bd1), objets_fixés : set(objet_bd2), seuils : list(réel), pas[15] : réel) : set (n-uplet (objet_comparé : objet_bd1, objets_fixés : set(objet_bd2))

La signature des autres outils d'appariement est donnée en annexe 7.3. Cet ensemble d'outils de type distance est incomplet, il manque des outils de distance entre des objets de dimensions différentes (2.3.3). Cependant, il permet déjà d'apparier géométriquement des objets en s'appuyant sur leur localisation.

#### 5.1.1.3 Les outils géométriques de type forme

Les objets géographiques peuvent aussi être caractérisés par leur forme [McMaster 86], [Kidner 96] [Buttenfield 91] [Plazanet 96]. Contrairement aux outils de type distance, les outils géométriques de forme caractérisent uniquement des couples d'objets. Ils sont donc adaptés uniquement pour des appariements 1-1. Ces outils ont été décrits dans l'état de l'art (2.3.2.2.3). Ils ne seront donc pas détaillés à nouveau. Néanmoins, leur signature est donnée en annexe 7.3. Il existe un grand nombre de critères pour apparier les objets selon leur forme. Quelques études [Kidner 96], [Mustière 95] sur leur robustesse ont déjà été réalisées pour comparer des objets généralisés avec les objets avant généralisation. Pour aider l'utilisateur à choisir ces outils géométriques de forme, il serait nécessaire de réaliser des études similaires dans le cadre de l'appariement.

#### 5.1.1.4 Les outils topologiques

Les outils topologiques sont plus difficiles à définir, ils sont employés pour tous les types d'appariements (1-1, 1-n et n-m) et sont toujours utilisés en complément d'un autre outil d'appariement.

Pour une topologie **de graphe**, l'outil du **plus court chemin** est nécessaire. Les algorithmes de plus court chemin ([Zhan 96], [Dijkstra 59],…) sont généralement utilisés dans le domaine du transport pour déterminer l'ensemble des tronçons formant le plus court chemin entre deux points d'un graphe. Ce graphe peut être orienté ou non orienté. Ces algorithmes sont détournés et utilisés comme des **filtres** pour supprimer, dans un sous-graphe, les éléments inutiles pour aller d'un point à un autre. Pour des éléments routiers, l'algorithme doit prendre en compte le **graphe de communication** (communication restreinte par les sens uniques et les interdictions venant d'un tronçon « t1 » de tourner vers un tronçon « t2 » pour le noeud « n »). Un algorithme de plus court chemin tenant compte du graphe de communication a été développé sous $GéO_2$. Il est décrit dans [Areia 96]. Il prend en paramètre un sommet de départ, un sommet d'arrivée et un graphe (incluant les sommets de départ et d'arrivée). Il renvoie l'ensemble d'arêtes formant le plus court chemin entre les sommets de départ et les sommets d'arrivée. Le graphe peut être orienté ou non.

[15] un pas de rasterisation est nécessaire car cet outil rasterise les lignes.

**Plus_court_chemin** (sommet_départ : objet_bd2, sommet_arrivé : objet_bd2, graphe : set (objet_bd2)) : set (objet_bd2),

Dans le prototype, d'autres outils topologiques plus simples ont aussi été implementés. Ce ne sont pas des outils d'appariement à proprement parler (une seule base est utilisée). Cependant, ils servent à contrôler des propriétés que doivent vérifier les sommets et les arêtes des réseaux à apparier, et donc à affiner les appariements.

Le premier outil de ce type est le **regroupement connexe**. Les n éléments intervenant dans un appariement 1-n doivent former un ensemble cohérent (un chemin, un carrefour complexe). Ces ensembles se caractérisent par la connexité de leurs éléments. Il est donc nécessaire de disposer d'un outil formant des groupes connexes à partir d'un ensemble de sommets et d'arêtes.

**Regroupement_connexe** (arêtes : set (objet_bd), noeuds : set (objet_bd)) : set(tuple (arêtes_du groupe : set (objet_bd), noeuds_du_groupe : set (objet_bd)))

Si ce processus renvoie un seul groupe connexe, cet ensemble est connexe.

Le deuxième outil **Impasse** sert à tester si le tronçon passé en paramètre est un cul-de-sac.

**Impasse** (**arête** : objet_bd1) : booléen

Le troisième, **nb_arêtes**, est utilisé pour renvoyer le nombre d'arêtes reliées au sommet passé en paramètre, prenant les valeurs $V_i$ pour les propriétés $P_i$ (attributs ou méthodes). Il peut servir, par exemple pour déterminer le nombre de tronçons reliés au sommet faisant partis du même carrefour complexe (une méthode testant si ce tronçon fait partie du carrefour complexe sera au préalabe définie).

**nb_arêtes** (sommet : objet_bd1, propriétés : list(string), types : list(string), valeurs : list(string)) : entier

Des outils servant à **prolonger** une arête pour atteindre un sommet doivent de même être inclus dans la boîte. Ils sont utilisés pour compléter un premier appariement. Des outils de ce type ont été utilisés entre autres par [Bucaille 96] [Phalakarn 91]. Ils prennent en entrée : l'arête à prolonger, le sommet à atteindre et le graphe des tronçons candidats. Par contre, le critère servant à sélectionner les arêtes peut varier (plus court chemin, angle, distance,…).

Pour une topologie de **carte**, des outils d'appariement topologique ont aussi été définis ([Phalakarn 91] [Servigne 93]) pour apparier des surfaces à partir de l'appariement des frontières et pour former des partitions de l'espace équivalentes.

#### 5.1.1.5 Les outils mixtes

Des outils mixtes peuvent être définis. Par exemple, un outil de **plus proche chemin** sera proposé comme extension pour le prototype routier. Il prend en paramètre un sommet de départ, un sommet d'arrivée, un graphe et des arêtes de l'autre base formant un chemin. Il renvoie l'ensemble des arêtes formant le chemin entre le sommet de départ et le sommet d'arrivée, et dont la distance moyenne[16] au chemin de l'autre base est la plus faible (le graphe peut être orienté ou non).

**Plus_proche_chemin** (sommet_départ : objet_bd2, sommets_arrivé : objet_bd2, graphe : set (objet_bd2), chemin : set(objet_bd1)) : set (objet_bd2),

[16] Une autre distance peut aussi être employée

Cet ensemble d'outils d'appariement peut être complété à n'importe quel moment par des outils atomiques ou des macro-outils (c'est-à-dire définis à partir des outils de la boîte). Une fois cette boîte établie, les différentes étapes du processus générique peuvent être présentées.

## *5.1.2 Les étapes du processus générique*

Les étapes d'appariement présentées dans cette section, s'appuient sur les expériences d'appariement complexes réalisées au laboratoire COGIT :

- L'appariement de BD surfaciques (appariement surface-surface et surface-ligne) pour comparer la cohérence géométrique du Cadastre et de la BD TOPO [Lemarié 96]. Cet appariement a démontré que l'on pouvait aussi apparier des objets surfaciques.
- L'appariement de BDG routières à différentes échelles [Devogele et al. 96 a] [Devogele et al. 96 b] décrit dans le chapitre suivant. Cet appariement a montré qu'un processus d'appariement complet de BDG devait être décomposé en séries d'appariement à effectuer dans un ordre précis.
- L'appariement de BDG routières selon différents points de vue à la même échelle [Branly 97] qui a révélé que des appariements provisoires pouvaient être repris et améliorés lors de la réalisation d'autres appariements.
- L'appariement de BDG ayant la même représentation mais à différentes dates [Bucaille 97]. Cette expérience a établi que plusieurs outils d'appariement devaient être lancés parallèlement, chacun des outils étant insuffisant et ne donnant de bons résultats que pour une partie des données. En outre, une deuxième étape consistant à fusionner les résultats des différents appariements, doit être réalisée afin de conserver la meilleure part de chaque appariement.

Donc, un processus générique d'appariement doit permettre d'utiliser conjointement plusieurs outils en parallèle ou en série, il se décompose en plusieurs phases.

### 5.1.2.1 L'enrichissement des BDG

La première étape est l'enrichissement des BDG par des **attributs virtuels** (méthodes) et des caractéristiques sur la forme de l'objet (aire, distance radiale au centroïde, …). Cette phase permet la **mise en conformité** des informations et la **matérialisation** d'informations implicites. L'utilisateur dispose alors des données requises par les outils d'appariement.

### 5.1.2.2 La sélection d'objets candidats à l'appariement

La deuxième étape est la sélection : elle désigne dans les deux bases, les groupes d'objets candidats à l'appariement. La sélection des objets candidats permet de restreindre la recherche à une population ciblée afin de limiter le nombre d'éléments parasites et les temps de calcul. Ces sélections peuvent être :

- des populations de classes entières,
- deux ensembles d'objets répondant aux mêmes critères (ce critère peut être un appariement déjà réalisé),
- une instance d'une classe de la première base et les instances candidates à l'appariement de la deuxième base en fonction de propriétés de la première instance.

L'étape de sélection s'appuie donc sur des **outils de sélection** qui sont des outils d'appariement rudimentaires (rectangle englobant, distance euclidienne, …) et des **appariements déjà réalisés**.

De plus, un **ordre** sur les sélections doit être établi pour enchaîner les appariements. Les appariements les plus fiables et qui interviennent lors d'autres appariements doivent êtres exécutés les premiers. Cet ordre va conditionner la qualité du processus global.

Il faut aussi noter qu'un élément candidat à une sélection, qui n'a pas été apparié, peut être sélectionné à nouveau.

#### 5.1.2.3 Les calculs de mesures d'appariement

Après avoir effectué une sélection, des mesures d'appariement entre des objets sélectionnés ou des objets en relation avec ces derniers sont calculées.

La difficulté de cette étape est le **choix** des outils à utiliser et de leurs paramètres. Tout d'abord, l'utilisateur doit rechercher s'il existe des **identifiants communs** (Numéro INSEE, …) aux deux bases. Dans l'affirmative, il est souhaitable d'utiliser l'outil d'égalité sémantique. Sinon, l'utilisateur doit utiliser des outils géométriques et topologiques. Le choix de l'outil se fait alors en fonction des critères suivants :

- le **type de l'appariement** recherché (appariement 1-1, 1-n et n-m),
- la **proximité** des objets correspondants,
- la **similitude des formes** des objets correspondants,
- les **relations topologiques** entre les objets à apparier et les objets déjà appariés.

Selon les réponses apportées à ces questions, l'utilisateur choisira un ou plusieurs des outils répondant à ces critères.

#### 5.1.2.4 Le filtrage et la prolongation

Cette étape permet d'interpréter les résultats de la phase précédente et de faire évoluer les ensembles d'objets candidats à l'appariement.

Le **filtrage** consiste à supprimer les objets parasites parmi les éléments candidats. Ces objets sont détectés soit grâce aux mesures réalisées lors de la phase précédente, soit à l'aide de nouveaux outils (plus court chemin, …). Ainsi, les objets ne répondant pas aux **critères** fixés par l'utilisateur sont supprimés. Le choix des valeurs de ces critères est réalisé, soit à l'aide de méta-données, soit empiriquement par tâtonnement.

Le filtrage peut aussi être utilisé pour **affiner des appariements déjà réalisés**. En effet, un processus d'appariement (appelé **appariement provisoire**) peut être repris une fois qu'un autre processus aura été effectué. Cette consolidation d'appariement provisoire par d'autres appariements est très utile pour rendre fiable l'appariement des objets en relation. Cette méthode a été utilisée dans [Branly 97] pour apparier les tronçons et les noeuds de la BD TOPO et de GEOROUTE.

La **prolongation** consiste à compléter une sélection si celle-ci, aux vues des mesures d'appariement, est jugée incomplète pour en déduire un appariement fiable. Pour ces nouveaux objets sélectionnés, la phase de mesure d'appariement doit être lancée. Par exemple, pour apparier les tronçons de réseaux, si les mesures réalisées indiquent que la sélection est incomplète, le plus petit des deux chemins est prolongé en ajoutant une des arêtes suivantes, afin d'établir des liens n-m entre les deux bases.

#### 5.1.2.5 Le regroupement d'appariements

L'utilisation d'outils d'appariement 1-1 pour des appariements de type 1-n ou l'utilisation d'outils d'appariement 1-n pour des appariements de type n-m crée des appariements

incomplets. Il est donc nécessaire de regrouper les résultats obtenus pour retrouver les appariements entre objets. Cette phase est simple et consiste à confronter les différents regroupements obtenus, afin de détecter les objets apparaissant dans plusieurs groupes. Elle est réalisée une fois que tous les objets susceptibles d'être regroupés sont appariés provisoirement.

#### 5.1.2.6 L'analyse du résultat, le contrôle de cohérence

Durant les phases précédentes, certains **contrôles de cohérence** ont pu déjà être réalisés implicitement lors de l'utilisation des outils. Par exemple, l'utilisation d'un algorithme de plus court chemin permet de filtrer les sélections, mais aussi de contrôler la connexité du chemin. Cependant, toutes les contraintes de cohérence n'ont pas forcement été vérifiées. Il faut donc avant de valider les résultats obtenus, analyser les correspondances afin de vérifier la validité des contraintes non encore employées [Ousset 97]. Ces contraintes peuvent porter sur la cardinalité des appariements obtenus, la connexité, la complétude, le graphe de communication, les relations de composition, ….

Si, les objets vérifient les contraintes déterminées, l'appariement pourra être considéré comme valide. Par contre, si les objets en correspondance ne les vérifient pas un **contrôle de cohérence manuel** devra être appliqué.

#### 5.1.2.7 Enchaînement des phases

Une fois toutes ces phases décrites, il est nécessaire de définir leur enchaînement (figure 60).

La phase d'enrichissement est la première phase, elle est exécutée une seule fois au début du processus, tandis que, les autres phases sont exécutées en boucle.

En premier lieu, une sélection est réalisée sur les deux bases à apparier.

Puis, des mesures d'appariement sont réalisées sur les objets sélectionnés (les objets candidats à l'appariement).

Pour une stratégie par **filtrage**, en fonction de ces mesures et des appariements (provisoires ou consolidés) déjà réalisés, une partie des objets candidats sont supprimés. Ce filtrage et ces mesures peuvent aussi servir à filtrer à nouveau des appariements provisoires déjà établis.

Pour une stratégie par **prolongation**, si les mesures d'appariement et les appariements (provisoires ou consolidés) montrent que les ensembles d'objets candidats sont incomplets, d'autres objets seront sélectionnés et de nouvelles mesures seront exécutées. Cette étape de filtrage ou de prolongation donne des appariements provisoires.

Une fois ces appariements provisoires réalisés, d'autres objets des deux bases peuvent être sélectionnés. Quand tous les objets faisant partie des classes impliquées dans une même ACI (Assertion de Correspondance Interschémas) ont été appariés provisoirement, les différents appariements sont confrontés pour regrouper si nécessaire, les appariements provisoires 1-1 ou 1-n.

Finalement, les appariements obtenus sont analysés pour vérifier les contraintes d'appariement définies qui n'ont pas été utilisées lors des phases précédentes. Trois issues sont alors possibles :

- l'appariement est valide,
- l'appariement est incohérent et un contrôle de cohérence manuel des données doit être réalisé,
- les objets sélectionnés ne s'apparient pas.

Cette analyse étant faite, de nouvelles sélections ont lieu pour les objets des classes des autres ACI, jusqu'à ce que le processus d'appariement ait sélectionné tous les éléments susceptibles d'être appariés.

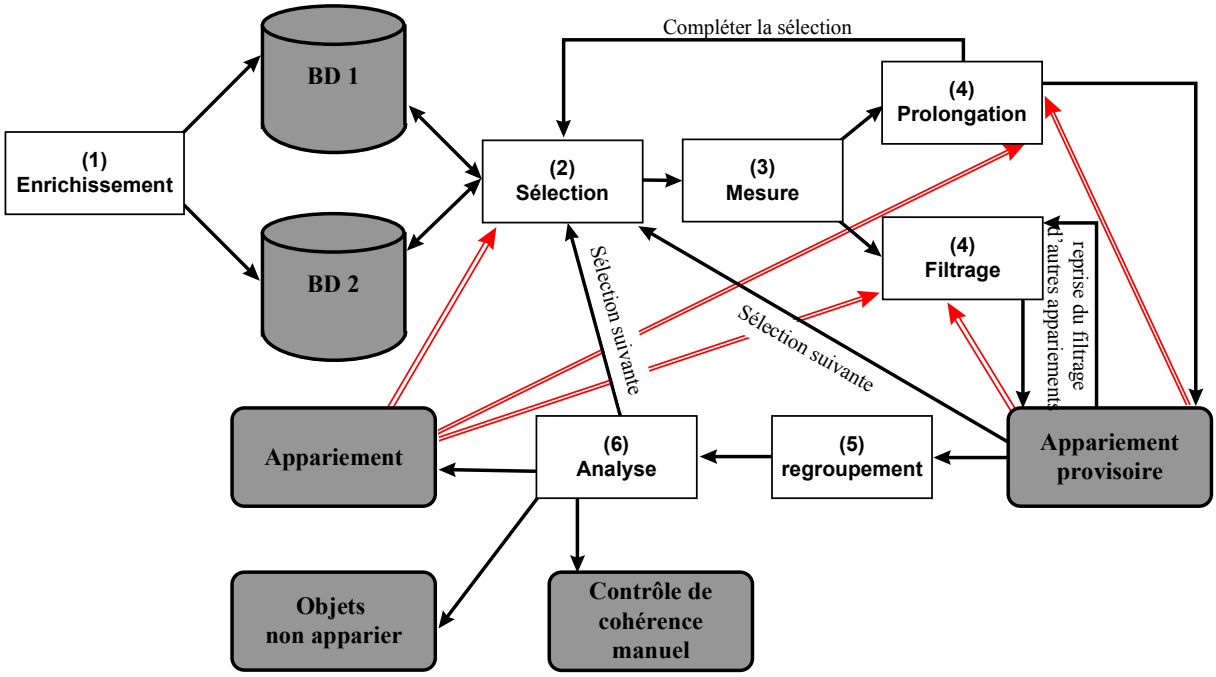



figure 60 : Enchaînement des phases du processus générique d'appariement

### *5.1.3 Conclusion sur le processus générique et la boîte à outils*

A partir de l'analyse des différents travaux d'appariement réalisés au COGIT, un processus générique a été défini. Il s'appuie sur une boîte à outils d'appariement qui regroupe l'ensemble des outils d'appariement disponibles. Le processus générique définit les **phases communes** à tout processus d'appariement, quelles que soient les données à apparier et leur **enchaînement**. La définition d'une méthode générique est un premier pas fondamental pour l'appariement mais il n'est pas suffisant. En effet, l'utilisateur doit encore déterminer l'ordre des sélections, choisir les outils d'appariement puis leurs paramètres, et enfin opter pour une stratégie de filtrage ou de prolongation. Le processus générique doit donc être complété par des méthodes d'aide au choix des outils, de leurs paramètres et de leur enchaînement.

Ce processus d'appariement doit aussi être rendu moins déterministe. Les processus définis apparient les données si un ensemble de critères d'appariement est vérifié. Or, visuellement, nous pouvons accepter des appariements ayant un critère non vérifié si les autres critères d'appariement sont vérifiés. Des mécanismes utilisant une logique floue [Bouchon-Meunier 94], répondraient à cette requête et devraient être étudiés d'une manière approfondie (une thèse doit commencer sur ce sujet au laboratoire COGIT).

## 5.2 Processus d'appariement de BD routières à différentes échelles

Le prototype mis en place pour apparier les instances de GEOROUTE et de la BD CARTO sur la zone de Marne-la-Vallée pour le thème routier va maintenant être exposé. Il a permis de relier automatiquement les instances de ces deux BDG et de contrôler la cohérence des deux représentations. Cette description permettra aussi, d'illustrer le processus d'appariement générique.

Les données routières sont le plus souvent construites selon une structure de graphe, autour de trois concepts clés[17] :

- le **noeud routier** est un sommet du graphe routier, il peut représenter, selon l'échelle, un cul de sac, un carrefour simple, un rond-point, un échangeur, voire une ville,
- le **tronçon routier** est une arête du graphe routier, il représente une portion de route de valeur homogène pour ces attributs, située entre deux noeuds routiers (ses extrémités),
- la **route** est un ensemble de tronçons routiers.

C'est l'appariement de ces trois concepts qui va être décrit maintenant.

Comme nous pouvons le constater par la comparaison des deux figures suivantes (figure 61 et figure 62), les représentations des mêmes phénomènes du monde réel dans ces deux bases sont très différentes :

- A une instance de la classe ROUTE de la BD CARTO correspond une instance de la classe ROUTE de GEOROUTE.
- A une instance de la classe TRONÇON de la BD CARTO correspond une ou plusieurs instances de la classe TRONÇON de GEOROUTE.
- A une instance de la classe NOEUD de la BD CARTO peut correspondre une ou plusieurs instances de la classe NOEUD et zéro à plusieurs instances de la classe TRONÇON de GEOROUTE.
- Certaines instances des classes TRONÇON et NOEUD de GEOROUTE n'ont pas forcement de correspondant.

[17] D'autres phénomènes du monde réel sont reliés à ces concepts comme les aires de péage, les aires de repos, les parkings,…

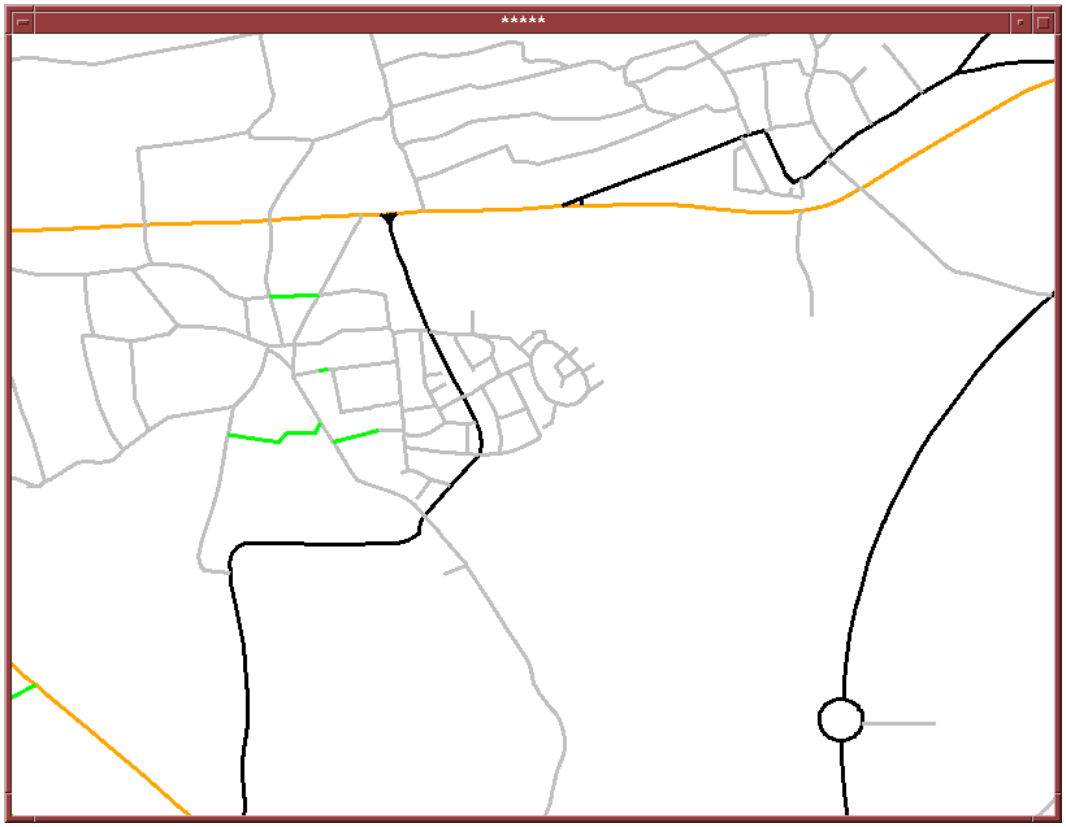



figure 61 : GEOROUTE Montévrain

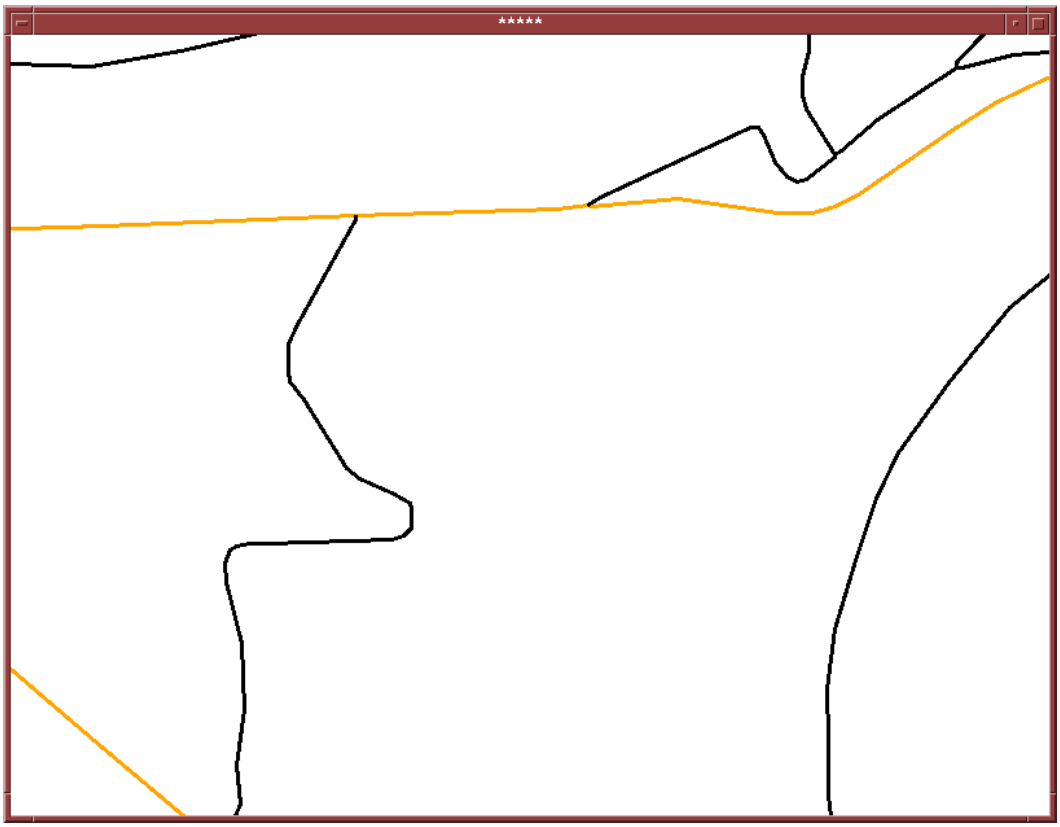



figure 62 : BD CARTO Montévrain

Ces informations (qui sont contenues dans les ACI) nous ont permis de distinguer trois sous-processus d'appariement :

- l'appariement des routes BD CARTO,
- l'appariement des noeuds BD CARTO,
- l'appariement des tronçons BD CARTO.

Ces sous-processus exploitent les trois types d'information (sémantique, géométrique et topologique). Chacun de ces sous-processus sera détaillé par la suite (appariement des routes en 5.2.2, des noeuds en 5.2.3, des tronçons en 5.2.4). Auparavant, l'enchaînement de ces trois appariements est présenté (0). En dernier lieu, une analyse des résultats obtenus sera réalisée (5.2.5).

### *5.2.1 Enchaînement des appariements*

Pour les données routières qui forment un réseau et sont donc forcément interconnectées, **l'ordre** des différents appariements est primordial.

Pour le prototype, l**'ordre** des sous-processus (figure 63) est le suivant. Les routes et les noeuds sont tout d'abord appariés. Les tronçons le sont en dernier lieu.

**L'appariement des routes** prend en entrée :

- les instance de la classes ROUTE de la BD CARTO,
- les instance de la classes ROUTE de GEOROUTE,

et génère des couples d'objets formés d'un objet de la classe ROUTE de la BD CARTO et d'un objet de la classe ROUTE de GEOROUTE.

**L'appariement des noeuds routiers** prend en entrée :

- les instances de la classe NOEUD de la BD CARTO,
- les instances de la classe NOEUD de GEOROUTE,
- les instances de la classe TRONÇON de GEOROUTE (pour former des carrefours complexes si nécessaire),

et génère des couples formés d'une instance de la classe NOEUD de la BD CARTO et d'un ensemble d'instances des classes NOEUD et TRONÇON de GEOROUTE.

**L'appariement des tronçons** est traité par la suite, puisqu'il utilise les résultats des deux appariements précédents. Il prend en entrée :

- les instances de la classe TRONÇON de la BD CARTO,
- les instances de la classe TRONÇON de GEOROUTE qui n'ont pas été appariées au préalable,
- les résultats de l'appariement des routes,
- les résultats de l'appariement des noeuds.

Cet appariement est réalisé en deux temps. Dans un premier temps, les tronçons appartenant à des routes appariées sont traités route par route. Dans un deuxième temps, les tronçons restant sont appariés. L'appariement des noeuds et l'appariement des routes servent donc de filtres et d'enrichissement pour celui des tronçons.

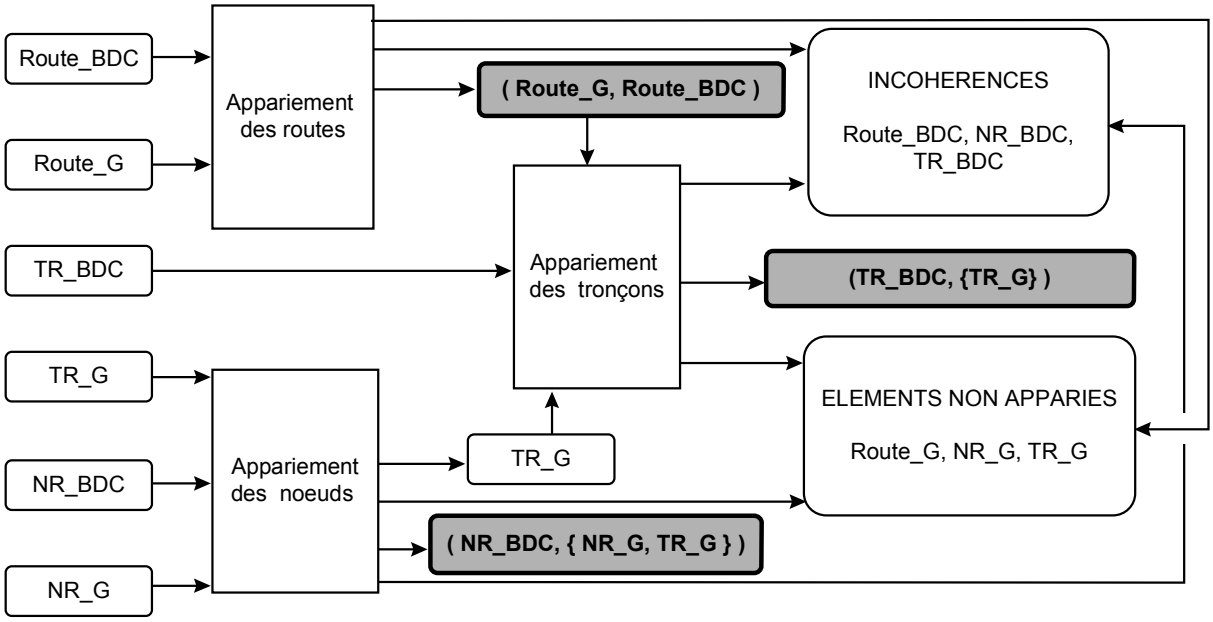



figure 63 : Processus d'appariement global du prototype

Les phénomènes du monde réel secondaires représentés dans GEOROUTE (tronçons de route communale, …) ne sont pas toujours représentés dans la BD CARTO. Les éléments de GEOROUTE non appariés sont donc principalement des phénomènes non sélectionnés dans la BD CARTO. Par contre, les instances de la BD CARTO non appariées représentent des phénomènes du monde réel dont les deux représentations sont incohérentes. **Une mise en conformité manuelle** (contrôle de cohérence) qui consiste à corriger les données de la représentation erronée ou à y ajouter des données, est réalisée pour ces derniers.

### *5.2.2 Appariement des routes*

L'appariement des routes est le plus simple des trois sous-processus. Les routes sont des ensembles de tronçons définis par un organisme (ministère de l'équipement,…). Elles possèdent un identifiant défini par cet organisme. L'appariement repose donc uniquement sur **l'égalité sémantique**. En effet, une clé d'identification peut être construite à partir des attributs *numéro* et *gestionnaire* de la BD CARTO et de GEOROUTE. Le numéro est une chaîne de caractères indiquant le numéro attribué à la route (N24, D134) et le gestionnaire est un attribut uniquement renseigné pour les routes départementales, il contient le numéro du département.

Par exemple, la route de la BD CARTO ayant pour valeurs « D134 » et « 77 » est appariée avec l'instance de GEOROUTE ayant pour valeurs « D134 » et « 77 », de même l'instance de la BD CARTO ayant pour valeurs « A4 » et « » est appariée avec l'instance de GEOROUTE ayant pour valeurs « A4 » et « ».

Cependant, avant d'apparier les routes, il est nécessaire de mettre les numéros de routes en conformité (retirer les espaces et mettre toutes les lettres en majuscule) grâce à des attributs virtuels. Ainsi, la «D34 a» pourra être appariée avec la «d 34A». Cette définition d'attributs virtuels est un exemple **d'enrichissement**.

La **sélection** des routes est effectuée en un seul passage (sélection de la population des deux classes). L'étape de **mesure** consiste alors juste à appliquer l'égalité sémantique. Les étapes suivantes (filtrage et regroupement) ne sont ici pas nécessaires.

Ce type de processus d'appariement est également réalisable pour des objets composés (département, rue, …) qui ont des identifiants définis par un organisme, il est similaire au mécanisme de la clause AIC (« Avec Identifiants Correspondants ») (3.1.4.2).

### *5.2.3 Appariement des noeuds routiers*

L'appariement des noeuds routiers est le plus complexe, car il repose sur des informations géométriques, topologiques et sémantiques. Il doit faire correspondre un noeud BD CARTO à un noeud GEOROUTE ou à un ensemble de noeuds et de tronçons de GEOROUTE formant un **carrefour complexe** : échangeur, rond-point, patte d'oie,… (figure 64).

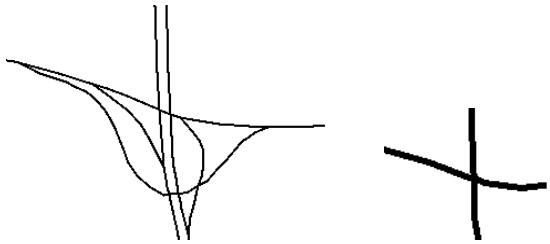


figure 64: Le même carrefour dans GEOROUTE et dans la BD CARTO

Sachant que le plus souvent un noeud de la BD CARTO correspond à un noeud de GEOROUTE, cette hypothèse sera vérifiée la première. S'il n'est pas possible de réaliser un appariement 1-1, l'appariement entre le noeud de la BD CARTO et un ensemble de noeuds et de tronçons de GEOROUTE sera alors essayé.

Pour chaque noeud de la BD CARTO, l'appariement se décompose donc en cinq étapes :

- la **sélection** d'un noeud de la BD CARTO et des noeuds GEOROUTE candidats (5.2.3.1),
- le calcul de **mesures** d'appariement (5.2.3.2),
- la **prolongation** par la sélection des tronçons GEOROUTE candidats si nécessaire (5.2.3.3)
- le **filtrage** des instances candidates parasites si nécessaire (5.2.3.4),
- l'**analyse** du résultat (5.2.3.5).

#### 5.2.3.1 Les étapes de sélections des objets candidats à l'appariement

Pour chaque noeud de la BD CARTO, la sélection est composée de 3 phases (choix d'une zone de recherche, recherche des noeuds de la BD CARTO en concurrence, sélection des noeuds candidats).

##### *5.2.3.1.1 Choisir le rayon de la zone de recherche*

La première phase consiste à déterminer une **zone de recherche** pour chaque instance de la classe NOEUD de la BD CARTO. Les noeuds de GEOROUTE sont ensuite recherchés dans cette zone. Comme il n'y a aucune raison de privilégier une direction, une zone circulaire a été choisie, elle est centrée sur l'instance de la BD CARTO. Le **rayon** est déterminé empiriquement en fonction de la valeur de l'attribut *type* de la BD CARTO :

*type* = « Changement d'attribut » Rayon = 50 m
*type* = « Carrefour simple ou rond-point » Rayon = 125 m
*type* = « Grand rond-point » Rayon = 300 m
*type* = « Echangeur complet ou partiel » Rayon = 450 m

Le rayon de la zone de recherche varie en fonction de l'emprise du phénomène du monde réel correspondant (plus le type du noeud est « complexe », plus l'emprise du phénomène du monde réel est grande). Pour un type de noeud « complexe », les objets homologues de GEOROUTE sont donc plus dispersés que pour un type simple.

Ce rayon devrait être normalement fonction de l'erreur maximale de la base la moins précise, mais cette méta-donnée n'est pas présente dans la BD CARTO.

#### *5.2.3.1.2 Rechercher les noeuds de la BD CARTO en concurrence*

Cette zone de recherche a été déterminée indépendamment des autres instances de la BD CARTO. Si les deux noeuds sont à une distance inférieure à la somme de leur rayon, les zones de recherche s'intersectent et des noeuds candidats peuvent être communs. Or, un noeud GEOROUTE ne peut être apparié qu'avec un seul noeud de la BD CARTO.

Pour résoudre ce problème, deux solutions sont envisageables. La plus simple consiste à réduire la zone de recherche à l'intersection du cercle et de sa cellule du **diagramme de Voronoï** (figure 65).

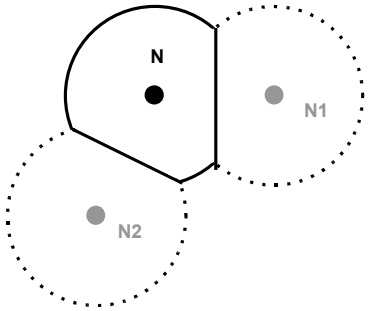



figure 65 : Zone de recherche réduite

Cette solution a été utilisée dans un premier temps. Le problème est que, certains noeuds de GEOROUTE devant être appariés avec un noeud BD CARTO sont plus proches d'un autre noeud de la BD CARTO. Cette méthode a priori a donc été remplacée par une méthode a posteriori qui consiste à détecter les appariements n-m durant la phase d'analyse (si deux noeuds BD CARTO sont appariés avec deux ensembles d'objets ayant une intersection non vide, ces deux appariements sont incohérents avec l'ACI).

#### *5.2.3.1.3 Sélection des noeuds GEOROUTE candidats*

Les noeuds de GEOROUTE se situant dans la zone de recherche sont sélectionnés, ils sont appelés **noeuds candidats**. Seuls les noeuds candidats pourront être appariés avec le noeud de la BD CARTO.

### 5.2.3.2 L'étape de mesure

L'étape de sélection a fourni des noeuds candidats. Il faut alors utiliser des outils d'appariement, afin de caractériser les éléments candidats. Comme les appariements de noeuds peuvent être de type 1-n sans processus d'appariement sémantique possibles, nous

utilisons les correspondances entre les tronçons qu'ils relient pour caractériser les noeuds candidats. Un appariement géométrique des tronçons reliés, encore appelé **pré-appariement géométrique des tronçons,** est donc réalisé.
L'outil sélectionné est l'appariement géométrique des tronçons de Stricher amélioré. Il est appliqué aux tronçons communicants[18] du noeud de la BD CARTO et des noeuds de GEOROUTE. Les seuils choisis sont de 30 mètres, 20 mètres et 10 mètres (ces seuils ont été également déterminés empiriquement).

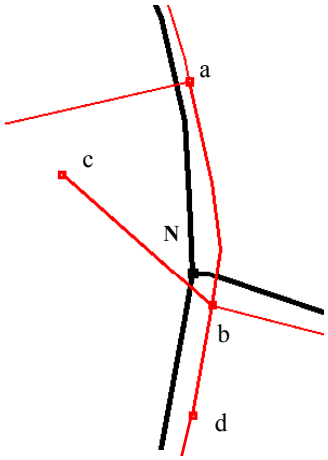



figure 66 : Appariement géométrique des tronçons communicants des noeuds candidats

Un noeud candidat à l'appariement avec le noeud de la BD CARTO (« nc ») peut être de deux types :

- **complet,** chaque tronçon communicant de « nc » s'apparie géométriquement avec au moins un tronçon communicant de ce noeud,
- **incomplet** sinon.

Ainsi, pour la figure 66, parmi les noeuds candidats de GEOROUTE (« a », « b », « c », « d »), seul le noeud « b » est complet, car trois de ses tronçons communicants sont appariés géométriquement avec les trois tronçons communicants du noeud « N » le noeud de la BD CARTO.

Cet appariement doit tenir compte du **sens de circulation** des tronçons. En effet, si un tronçon de la BD CARTO à double sens s'apparie uniquement avec un tronçon de GEOROUTE à sens unique, cet appariement n'est pas suffisant. Par contre, si le tronçon de la BD CARTO à double sens s'apparie avec deux tronçons de GEOROUTE en sens unique mais de sens opposé, l'appariement peut être accepté.

Trois cas sont possibles :

- il existe un noeud complet, alors seul ce noeud est conservé et nous passons directement à l'étape d'analyse (5.2.3.5) pour valider cet appariement 1-1,
- il existe plusieurs noeuds incomplets, nous passons alors à l'étape de prolongation,
- il existe un seul candidat qui est incomplet, un contrôle de cohérence manuel doit dans ce cas être réalisé.

Il faut noter que le cas : plusieurs noeuds complets, est impossible avec des données réelles et un seuil cohérent.

[18] un tronçon communicant d'un noeud est un tronçon relié à ce noeud par une relation NOEUD_INITIAL ou NOEUD_FINAL

#### 5.2.3.3 Les étapes de prolongation

Quand un appariement 1-1 est impossible, une étape de prolongation est définie afin de compléter la sélection, pour former des groupes de tronçons et de noeuds candidats.

##### *5.2.3.3.1 Sélection des tronçons GEOROUTE candidats*

Un **tronçon candidat** est un tronçon qui relie deux noeuds candidats par des instances des relations NOEUD_INITIAL ou NOEUD_FINAL (figure 67). Les noeuds et les tronçons candidats forment l'ensemble des éléments candidats à l'appariement.

##### *5.2.3.3.2 Formation des groupes candidats*

Une fois, les tronçons candidats sélectionnées, les **groupes connexes** sont formés, en utilisant les relations topologiques NOEUD_INITIAL ou NOEUD_FINAL entre les classes NOEUD et TRONÇON. Cette opération consiste à regrouper en sous-ensembles les éléments (noeuds, tronçons) candidats connexes. Pour la figure 67, quatre groupes connexes sont ainsi formés.

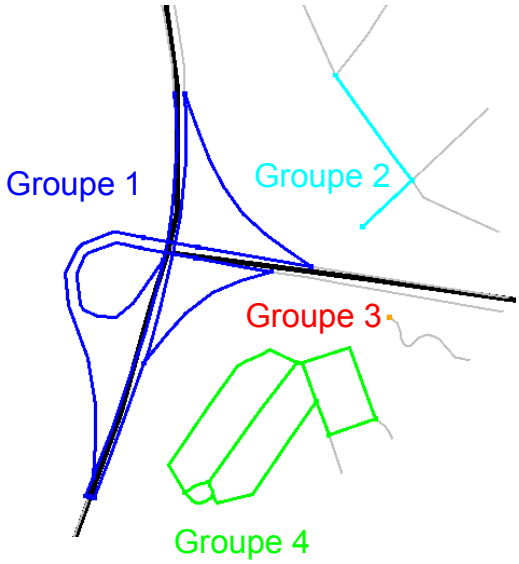



figure 67 : Formation des groupes connexes GEOROUTE pour
un noeud BD CARTO de type « échangeur complet »

Un tronçon d'un groupe candidat est appelé **tronçon composant** de ce groupe. Un tronçon lié à un noeud du groupe candidat, mais ne faisant pas partie de ce groupe est toujours appelé **tronçon communicant** de ce groupe candidat. Si le groupe est réduit à un noeud (groupe 3 de la figure 67), tous ces tronçons reliés sont des tronçons communicants.

Cette sélection est un exemple de sélection des instances candidates à l'appariement de la deuxième base en fonction de l'instance de la première base et d'outils de sélection (outils d'appariement rudimentaires).

La sélection de la prolongation a fourni des groupes candidats qui sont caractérisés à l'aide du **pré-appariement géométrique des tronçons** déjà réalisé, aucune nouvelle mesure n'est donc calculée.

Un groupe candidat à l'appariement avec un noeud de la BD CARTO (« nc ») peut être de trois types :

- **complet,** chaque tronçon communicant de « nc » s'apparie géométriquement avec au moins un tronçon sortant ou composant de ce groupe,

- **partiel,** si un sous-ensemble des tronçons communicants de « nc » s'apparient géométriquement avec les tronçons communicants ou composants de ce groupe,
- **impossible** sinon.

#### 5.2.3.4 Filtrage des éléments de GEOROUTE

Une fois les résultats du pré-appariement des tronçons connus, les groupes candidats peuvent être filtrés. Le filtrage est l'étape la plus complexe, elle se compose de six phases. Elle consiste à réduire l'ensemble des éléments sélectionnés, afin de ne conserver que les éléments à apparier.

##### *5.2.3.4.1 Choix du groupe candidat à apparier*

Un noeud de la BD CARTO est apparié avec un seul groupe connexe. Il faut donc choisir le groupe candidat. Cinq cas sont envisageables :

- Un **seul groupe est complet**, ce groupe est sélectionné. Sur l'exemple de figure 67, seuls les tronçons sortants et composants du groupe 1, s'apparient géométriquement avec les tronçons communicants du carrefour de la BD CARTO. Le groupe 1 est donc retenu.
- Un **seul groupe est partiel,** certains tronçons communicants du noeud BD CARTO ne sont pas appariés avec les tronçons communicants ou composants du groupe. Une partie du carrefour dans le groupe choisi est manquante. Un contrôle de cohérence manuel est obligatoire. Le groupe partiel servira de point de départ pour vérifier la cohérence entre les deux bases.
- **Plusieurs groupes sont partiels**. Le choix du groupe doit alors être traité manuellement après un contrôle de cohérence. Les groupes partiels serviront de points de départs pour vérifier la cohérence entre les deux bases.
- **Aucun groupe n'est partiel**. L'appariement devra aussi être traité manuellement lors du contrôle de cohérence.
- **Plusieurs groupes sont complets**. Ce cas n'a pas été rencontré et est improbable.

Pour la figure 67, le processus d'appariement pourrait s'arrêter là, mais la plupart du temps, il est nécessaire de filtrer le groupe candidat retenu afin d'éliminer les éléments parasites.

##### *5.2.3.4.2 Filtrage par suppression*

Le filtrage consiste à enlever les éléments parasites du groupe. Il se déroule en **quatre phases** :

1. le filtrage des « non carrefours »,
2. le filtrage par réduction,
3. le filtrage par un algorithme du plus court chemin,
4. le filtrage par réduction à nouveau.

###### *5.2.3.4.2.1 Filtrage des « non carrefours »*

La première phase consiste à supprimer les tronçons et les noeuds ne pouvant pas faire partie d'un ensemble d'éléments formant le carrefour complexe correspondant. Ces éléments sont les **impasses** et les tronçons t répondant aux prédicats suivants :

- « t » a une des extrémités « n » qui a pour seul tronçon composant « t »,
- « n » n'a pas de tronçons communicants appariés géométriquement.

Les noeuds qui ne sont plus reliés au reste du groupe connexe sont aussi supprimés.

Ainsi, pour l'exemple de la figure 68, les tronçons supprimés dans cette phase sont :

- « a » qui est une impasse,
- « b », une des extrémités de « b » n'a que « b » comme tronçon composant et ses tronçons communicants « j » et « k » ne sont pas appariés géométriquement.

*5.2.3.4.2.2 Filtrage par réduction*

La deuxième phase est **récursive,** elle consiste à supprimer les tronçons qui doivent être appariés avec un des tronçons communicants du noeud de la BD CARTO et non avec ce noeud. Un tronçon composant « t » de ce type répond aux prédicats suivants :

- « t » est relié avec un noeud « n » qui a pour seul tronçon composant « t »,
- le ou les tronçons communicants « $tc_i$ » de « n » appariés géométriquement sont appariés avec le même tronçon communicant du noeud BD CARTO « tbdc »,
- « t » est apparié géométriquement avec « tbdc ».

Pour la figure 68, les tronçons supprimés dans cette phase sont :

- « c », une de ses deux extrémités a pour seul tronçon composant « c », « f » le seul tronçon communicant apparié géométriquement de cette extrémité, est apparié avec « 2 » et « c » est apparié géométriquement avec « 2 »,
- « d » qui est supprimé pour les mêmes raisons,
- « e » qui est éliminé dans un deuxième temps, une de ses deux extrémités a pour seul tronçon composant « e », « c » (le seul tronçon communicant apparié géométriquement) et « e » sont appariés géométriquement avec « 2 ».

Le résultat de ce filtrage est la figure 69.

Néanmoins, les deux premières phases de filtrage par suppression ne sont pas toujours suffisantes, elles ne permettent pas en effet, de supprimer les tronçons parasites formant des cycles. Ainsi, l'application des deux premières phases du filtrage par suppression à l'ensemble des données de la figure 70 (a) a pour résultat la figure 70 (b).

*5.2.3.4.2.3 Filtrage par un algorithme du plus court chemin*

Pour supprimer ces boucles, la troisième phase utilise un algorithme de calcul du **plus court chemin**, tenant compte du graphe de communication. Cette phase commence par définir des points d'entrée et des points de sortie (en rouge sur la figure 70 (b)).

Les **points d'entrée** sont les noeuds du groupe candidat ayant un tronçon communicant apparié géométriquement permettant d'entrer dans le groupe candidat.

Les **points de sortie** sont les noeuds du groupe candidat ayant un tronçon communicant apparié géométriquement et permettant de sortir du groupe candidat.

Pour filtrer, un algorithme de plus court chemin qui respecte le graphe de communication, est appliqué entre tous les points d'entrée et tous les points de sortie, du noeud de la BD CARTO. L'ensemble des tronçons non parcourus par un des plus courts chemins, est supprimé. Ainsi, pour la figure 70 (b), la face en haut à gauche a pu être éliminée.

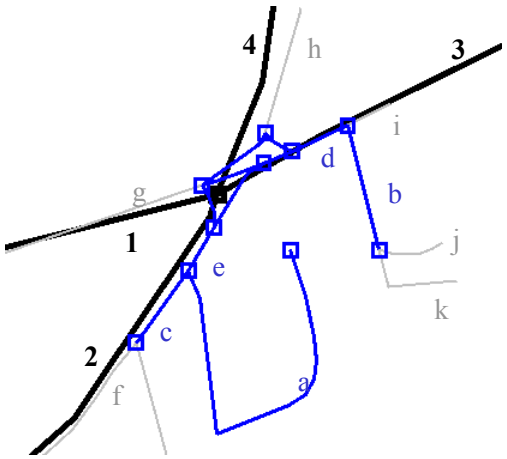



figure 68 : Groupe candidat avant le filtrage par suppression

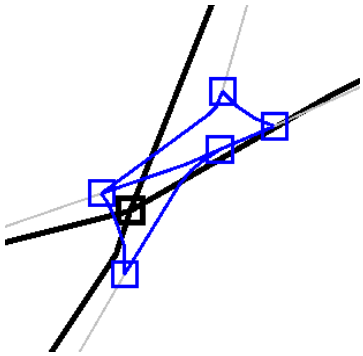
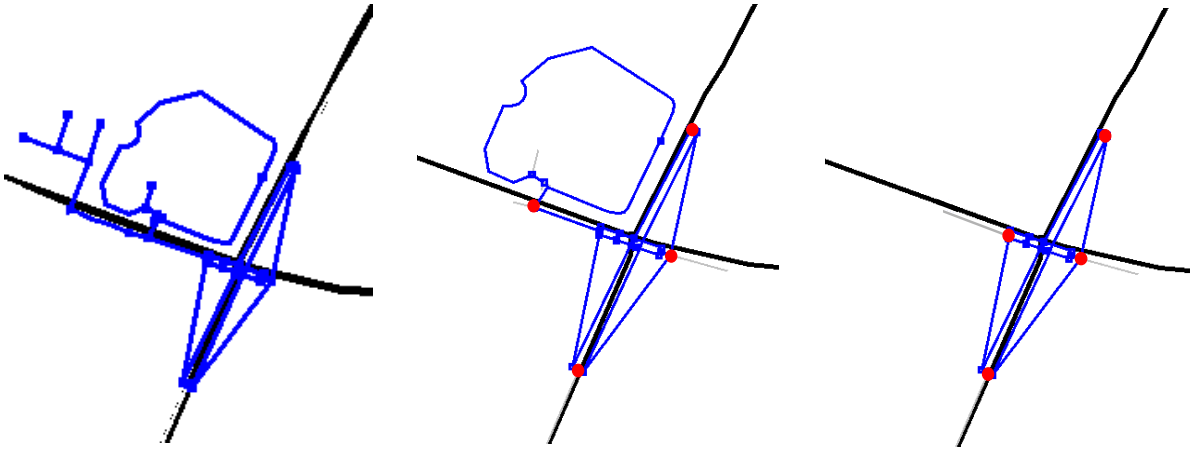


figure 69 : Groupe candidat après les deux premières phases du filtrage par suppression



(a) avant le filtrage

(b) après les deux premières phases du filtrage

(c) après le filtrage par suppression

figure 70 : Groupe candidat avant et après le filtrage

*5.2.3.4.2.4 Filtrage par réduction*

La quatrième et dernière phase du filtrage par suppression est strictement identique à la deuxième phase.

Ainsi, pour la figure 70 (b) le tronçon composant à droite de la face est ôté. Le résultat obtenu à la suite de toutes les phases du filtrage par suppression, pour la figure 70 (a) est visualisé dans la figure 70 (c).

### 5.2.3.5 Analyse

L'analyse du résultat consiste à vérifier les contraintes décrites dans l'ACI qui n'ont pas été vérifiées lors des précédentes phases. Durant la phase de filtrage, le graphe de communication du noeud de la BD CARTO a été vérifié pour les instances de GEOROUTE appariées avec ce noeud.

Par contre, les **appariements n-m** n'ont été détectés dans aucune phase. Il reste donc à contrôler dans cette phase, que tous les éléments de GEOROUTE apparaissent dans un seul appariement avec un noeud de la BD CARTO, afin de valider ceux-ci.

#### 5.2.3.6 Conclusion sur le processus d'appariement des noeuds routiers

Le processus d'appariement des noeuds routiers est complexe. La figure 71 résume l'ensemble des phases.

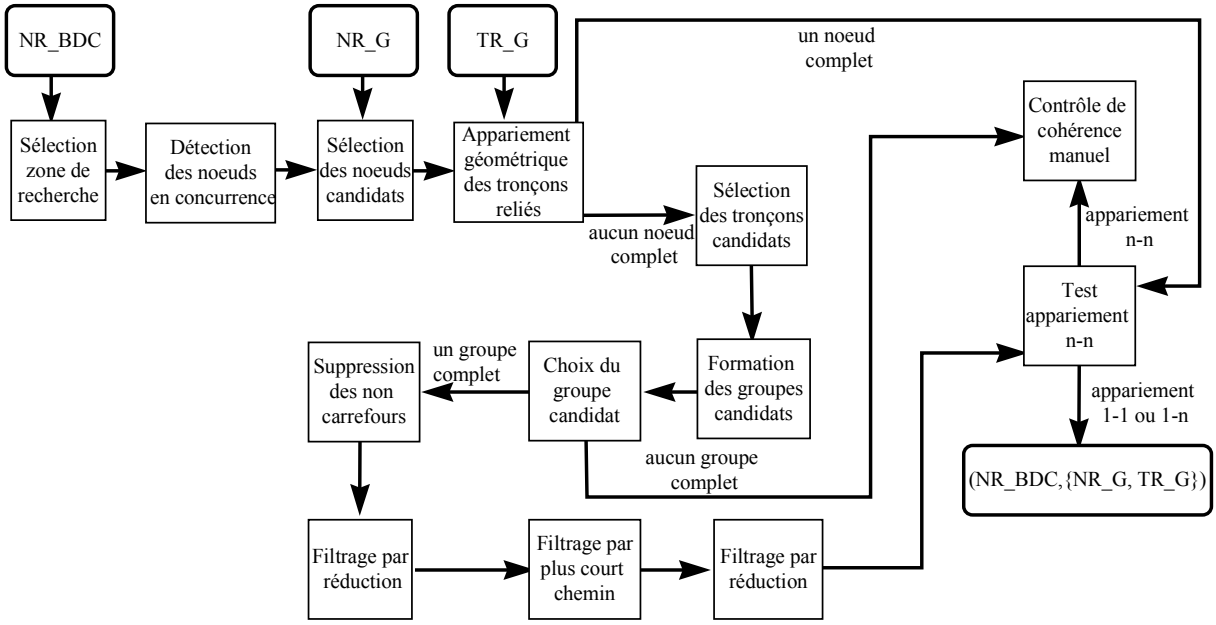



figure 71 : Les phases du processus d'appariement des noeuds de la BD CARTO

L'appariement des noeuds prend en compte plusieurs informations :

- la distance entre noeuds,
- la connexité des noeuds,
- un pré-appariement géométrique des tronçons,
- le plus court chemin entre les noeuds d'entrée et les noeuds de sortie du groupe connexe sélectionné.

Cette combinaison d'informations permet d'obtenir un appariement fiable (5.2.5) et de détecter les incohérences au niveau de l'existence des objets, de leur géométrie et de leurs relations topologiques.

### *5.2.4 Appariement des tronçons de route*

Le processus d'appariement des tronçons de route de la BD CARTO et des tronçons de routes de GEOROUTE est traité dans un deuxième temps, car il dépend des résultats obtenus pour les deux premiers appariements.

Il se décompose en quatre étapes :

- partition des classes TRONÇON de la BD CARTO et de GEOROUTE,
- calcul de mesures d'appariement (composante de la distance de Hausdorff),
- filtrage des tronçons de GEOROUTE candidats,
- analyse du résultat.

### 5.2.4.1 Partition des classes TRONÇON de la BD CARTO et de GEOROUTE

La partition des classes TRONÇON de la BD CARTO et de GEOROUTE est fonction des deux appariements précédents (figure 72). Une première sous-classe à part, est formée par les instances de la classe TRONÇON de GEOROUTE qui ont été appariées au préalable avec des noeuds. Ils ne seront donc pas appariés avec des tronçons. Puis, pour chaque couple de routes appariées, une sous-classe est créée dans chaque base ; elle regroupe les tronçons qui composent les routes appariées. Finalement, les tronçons ne faisant pas partie d'une route appariée forment les deux dernières sous-classes. Ainsi, l'appariement global de toutes les instances de la classe TRONÇON de la BD CARTO avec toutes les instances de la classe TRONÇON de GEOROUTE est scindé en plusieurs phases (un appariement sera réalisé pour chaque couple de sous-classes). Le nombre de tronçons parasites engendrés va pouvoir ainsi être diminué et le processus optimisé.

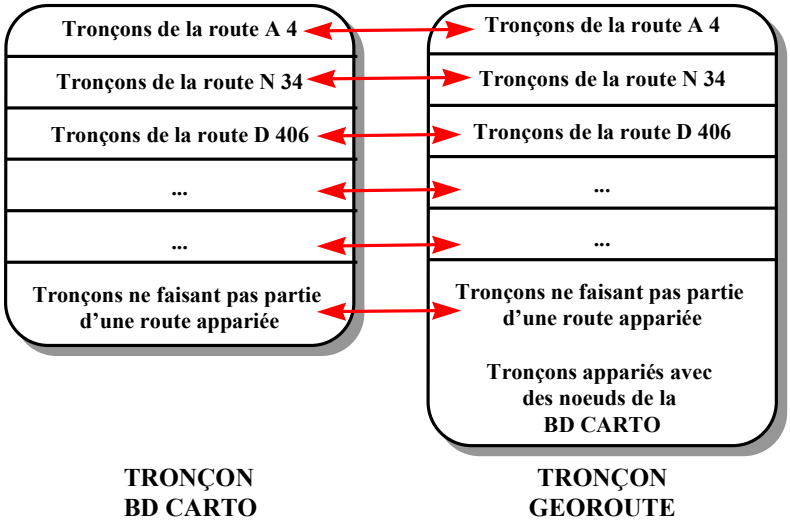



figure 72 : Partition des classes TRONÇON de la BD CARTO et de GEOROUTE

Cette partition est un exemple de sélection formant des ensembles répondant au même critère (l'appariement des routes). D'autre part, il illustre le fait qu'un élément déjà sélectionné non apparié peut être sélectionné à nouveau (un tronçon GEOROUTE candidat pour un noeud BD CARTO, non retenu, deviendra un tronçon candidat pour un tronçon). Finalement, il montre qu'à l'intérieur d'une classe un ordre peut être défini sur les instances (l'ordre d'appariement des tronçons est établi en fonction de l'appartenance à une route appariée).

### 5.2.4.2 Appariement géométrique

Une fois les sous-classes déterminées, l'appariement de leurs instances doit être réalisé. Dans cet objectif, l'outil d'appariement géométrique [Stricher 93] avec plusieurs seuils successifs, a été de nouveau utilisé. Les mêmes seuils de 30 mètres, 20 mètres puis 10 mètres pour les tronçons litigieux ont été employés.
Les tronçons de GEOROUTE sont donc de trois types : appariés géométriquement, litigieux, non apparié géométriquement.

Pour les tronçons non appariés géométriquement de GEOROUTE, deux sous-cas sont à distinguer :

- si le tronçon appartient à une route appariée, cette situation est **incohérente**, la composition des routes doit être contrôlée,

- sinon, le tronçon de GEOROUTE représente un phénomène du monde réel (tronçon du réseau secondaire, …) non représenté dans la BD CARTO.

Les tronçons appariés géométriquement sont sélectionnés, ils sont appelés **tronçons candidats**. L'appariement géométrique permet donc de sélectionner, pour chaque tronçon de la BD CARTO, un ensemble de tronçons GEOROUTE. La figure 73 donne un exemple des tronçons candidats (trait fin) renvoyés par cet appariement géométrique pour un tronçon de la BD CARTO (trait épais).

figure 73 : Exemple d'appariement géométrique à l'aide de la composante de Hausdorff

#### 5.2.4.3 Filtrage par un algorithme de plus court chemin et vérification de la connexité

Pour les tronçons de route appariée, le filtrage du résultat est inutile. Par contre pour les tronçons n'appartenant pas à une route appariée, il faut supprimer les parasites sélectionnés (impasses, chemins parallèles inutiles, …).

Pour filtrer l'appariement géométrique, trois propriétés ont été utilisées :

- Un tronçon « tc » de la BD CARTO sert à relier deux noeuds « a » et « b », l'ensemble des tronçons de GEOROUTE appariés avec ce tronçon doit donc permettre de relier les deux noeuds ou les deux carrefours complexes de GEOROUTE correspondants (cor(a), cor(b)). Si le tronçon « tc » est à double sens, il faudra établir dans GEOROUTE un chemin de cor(a) vers cor(b) et réciproquement. Par contre, si le tronçon « tc » est à sens unique de a vers b, il suffira d'établir un chemin de cor(a) vers cor(b).
- Ce ou ces chemins doivent être les plus « proches » possible du tronçon de la BD CARTO.
- Un tronçon de GEOROUTE doit être apparié avec un seul tronçon de la BD CARTO.

L'application de ces trois propriétés permet de supprimer les impasses et les chemins parallèles inutiles dans le sous graphe défini par l'ensemble des tronçons candidats. De plus, la connexité entre les noeuds ou les carrefours complexes correspondants peut être vérifiée.

Dans ce but, l'algorithme du plus court chemin respectant le graphe de communication [Areia 96] a été employé.

#### *5.2.4.3.1 Détermination des points de liaisons*

Avant de lancer l'algorithme du plus court chemin, il faut **déterminer les points de liaison** (points d'entrée et de sortie). Si un noeud de la BD CARTO correspond à un noeud « n » dans GEOROUTE, le noeud « n » est un point de liaison. En revanche, si le noeud de la BD CARTO correspond à un carrefour complexe, les points de liaison sont déterminés. Ces points sont les noeuds composants qui ont comme tronçon communicant un des tronçons candidats à l'appariement avec les tronçons de la BD CARTO. Pour connaître le type du point de liaison (point d'entrée ou point de sortie), il faut se baser sur le sens du tronçon candidat (tc) à l'appariement avec les tronçons de la BD CARTO. Ces points de liaison sont des **points d'entrée,** si « tc » est un tronçon à double sens ou à sens unique partant de ce point. En revanche, ces points de liaison sont des **points de sortie,** si « tc » est un tronçon à double sens ou à sens unique allant vers ce point.

Par exemple, dans le but d'apparier le tronçon « 1 » de la figure 74 trois points de liaison (entrée et sortie) sont déterminés « A », « B » et « C », car les tronçons candidats (en bleu) sont tous à double sens.

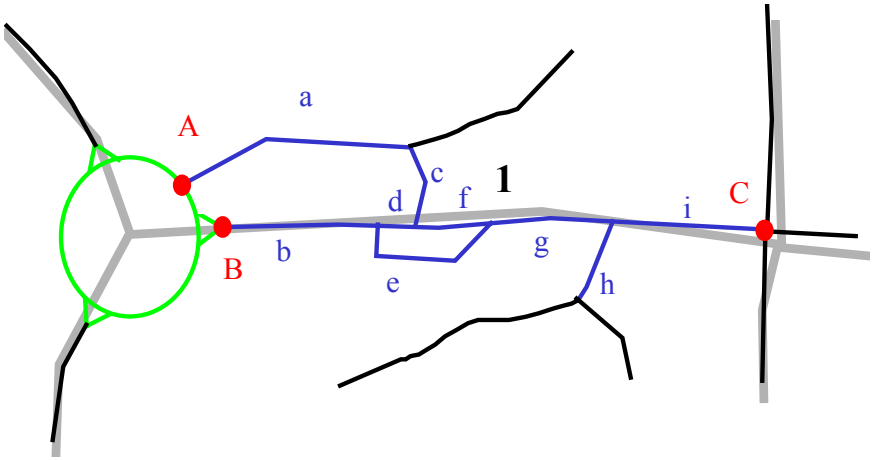



figure 74 : Exemple de points de liaison

#### *5.2.4.3.2 Calcul des plus courts chemins*

Une fois les points d'entrée et de sortie déterminés, les plus courts chemins peuvent être calculés. Le sens de communication du tronçon de la BD CARTO détermine les plus courts chemins à établir. Pour la figure 74, le tronçon de la BD CARTO étant à double sens, il faut calculer les plus courts chemins de « A » vers « C », de « B » vers « C », de « C » vers « A » et de « C » vers « B ». Les résultats sont les suivants :

- A → C  a, c, f, g, i
- B → C  b, d, f, g, i
- C → A  i, g, f, c, a
- C → B  i, g, f, d, b

#### *5.2.4.3.3 Choix des plus courts chemins*

Quand il existe un seul point de liaison dans chacun des carrefours complexes correspondants, il suffit de supprimer les tronçons n'apparaissant pas dans au moins un des deux chemins. La figure 75 donne le résultat obtenu une fois les tronçons candidats filtrés pour la figure 73.

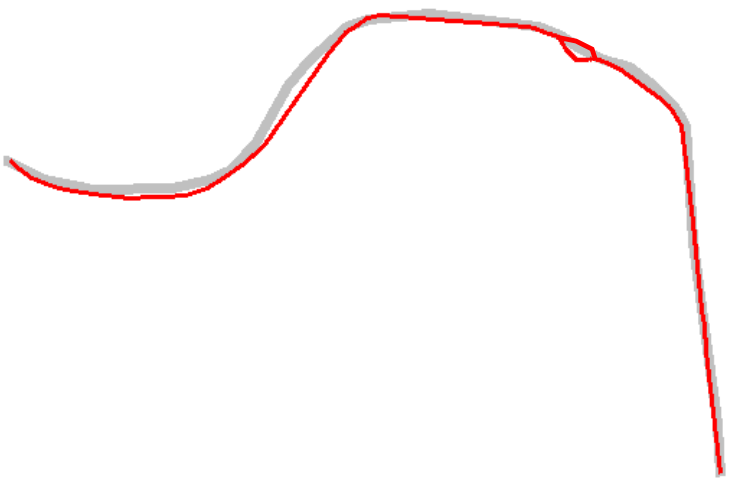


figure 75 : Exemple de filtrage par plus court chemin

Quand il existe plusieurs points de liaison dans au moins un des deux carrefours complexes ou noeuds correspondants, un seul chemin est retenu dans chaque sens. Par exemple, pour la figure 74, les chemins du point « A » vers le point « C » (A → C) et du point « C » vers le point « A » (C → A) sont des chemins parasites. Pour supprimer ces chemins parasites, il faut sélectionner **le plus court des plus courts chemins** allant dans le même sens. Ainsi, pour la figure 74, pour aller du carrefour complexe vers « C », il existe deux chemins (A → C et B → C), le chemin de « A » vers « C » est supprimé car le chemin de « B » vers « C » est plus court. De même, pour aller de « C » vers le carrefour complexe, il existe deux chemins (C → A et C → B), le chemin de « C » vers« A », plus long, est supprimé. Le tronçon « 1 » est donc apparié avec les tronçons b, d, f, g et i.

#### *5.2.4.3.4 Contrôle de cohérence*

Les phases précédentes ont déjà permis de vérifier :

- l'existence de tronçons homologues dans GEOROUTE pour chaque tronçon de la BD CARTO,
- la connexité des tronçons de GEOROUTE appariés (plus courts chemins) pour les tronçons n'appartenant pas à une route appariée.

Il reste donc à contrôler la connexité des tronçons de GEOROUTE appariés pour les tronçons appartenant à une route appariée,

Si cette contrainte n'est pas vérifiée, il faut lancer un contrôle de cohérence manuel.

### *5.2.5 Evaluation des résultats obtenus*

Le processus d'appariement développé sous $GéO_2$ dure environ 1 heure sur une SPARC 10.

Afin de le valider pour ces deux BDG, les résultats obtenus automatiquement pour la zone de Lagny Marne-la-Vallée ont été contrôlés visuellement (copies d'écran des résultat en annexe 7.5) .

#### 5.2.5.1 Evaluation de l'appariement des routes

Les résultats de **l'appariement des routes** sont de 100% d'appariement correct une fois les numéros de route mis en conformité. Par contre, les relations de composition (une route est

composée de tronçons routiers) sont trop souvent incohérentes, l'appariement des routes n'a donc pas été utilisé pour l'appariement des tronçons de route.

#### 5.2.5.2 Evaluation de l'appariement des noeuds

Les appariements automatiques des **343 noeuds routiers de la BD CARTO** ont été vérifiés visuellement, les résultats sont les suivants :

- Pour les **227 appariements 1-1** automatiques fournis par le processus,
    - 223 appariements sont corrects ,
    - 4 appariements sont incorrects,
        - 1 est incomplet (le tronçon GEOROUTE qui forme un rond-point en cul-de-sac avec le noeud GEOROUTE n'est pas sélectionné (annexe figure 96)),
        - 3 noeuds BD CARTO (cul-de-sac) sont appariés avec un noeud GEOROUTE erroné (noeuds GEOROUTE proches ayant un petit tronçon relié apparié avec le tronçon du noeud BD CARTO).
- Pour les **83 appariements 1-n** automatiques fournis par le processus,
    - 45 sont complets (les tronçons communicants de la BD CARTO sont appariés géométriquement) et sont donc filtrés,
        - 41 appariements filtrés sont corrects,
        - 4 incluent des parasites ou sont incomplets (problème de choix des points de communication),
    - 38 sont partiels (tous les tronçons communicants de la BD CARTO ne sont pas appariés), ils ne peuvent donc pas être filtrés par l'algorithme du plus court chemin,
        - 21 donnent le bon résultat mais seront détectés comme susceptibles d'être incohérents à cause des tronçons sortants non appariés géométriquement,
        - 17 sont incohérents.

    Les **causes d'appariement partiels** sont les suivantes :
    - des tronçons sortants manquants surtout en bordure de zone,
    - des défauts aux intersections de la BD CARTO [Bonin 95] (très petits tronçons dus au scannage et à la vectorisation (annexe figure 100)),
    - des défauts aux intersections de GEOROUTE discontinuité (annexe figure 99),
    - des décalages supérieurs à 30 mètres,
    - l'attribut type du tronçon de la BD CARTO a une valeur erronée.

    Ces 5 premières causes regroupent 31 cas qui sont de véritables incohérences,

    - des tronçons de la BD CARTO courts légèrement décalés par rapport aux tronçons GEOROUTE,
    - un échangeur dont les tronçons composants sortent de la zone de recherche (annexe figure 95),
    - des tronçons communicants GEOROUTE proches et parallèles (annexe figure 94).

    Ces 3 derniers cas sont des cas limites pour le sous-processus (7 cas).
- Pour les 4 **appariements n-m** (8 noeuds BD CARTO) détectés par confrontation des appariements 1-n complets et 1-1 :
    - 3 sont dus à des incohérences (représentation incohérente d'un carrefour complexe dans la BD CARTO, tronçon GEOROUTE non relié à son « vrai » noeud extrémité),

    - 1 est dû à 2 carrefours complexes se touchant (ce cas n'a pas été pris en compte par le processus).
- Pour les **25 non appariements** automatiques fournis par le processus, il y a :
    - 22 cas où il n'existe aucun noeud dans la zone de recherche,
    - 2 cas où il existe plusieurs groupes partiels (il manque des tronçons pour former le groupe connexe),
    - 1 cas où il existe un noeud dans la zone mais aucun tronçon communicant apparié.

Pour résumer, pour cette zone, le processus renvoie :
- 76,97 % d'appariements corrects,
- 2,33 % d'appariements incorrects (erreurs du processus),
- 12,54 % d'incohérences entre les représentations des carrefours du monde réel,
- 6,12 % d'incohérences dues à l'appariement géométrique des données en relation,
- 2,04 % d'incohérences dues aux limites du processus.

Pour les incohérences, il faut nuancer les résultats et distinguer **deux types d'incohérences** :
- les incohérences liées à l'absence d'un noeud ou d'un tronçon de GEOROUTE en bordure de zone qui représente environ la moitié des incohérences. Dans ce cas, l'appariement ne peut pas être réalisé ou validé mais la détection d'une incohérence ne va pas toujours engendrer une correction.
- dans les autres cas, la détection d'une incohérence entraîne la correction des données (noeuds BD CARTO, noeuds GEOROUTE, tronçons GEOROUTE composants ou communicants).

Visuellement, nous avons constaté qu'une correction devra être réalisée dans environ 5% des cas dans GEOROUTE et dans 2% des cas dans la BD CARTO.

#### 5.2.5.3 Evaluation de l'appariement des tronçons

Pour les **533 tronçons routiers de la BD CARTO**, les appariements automatiques ont été vérifiés visuellement, les résultats sont les suivants :

- **374 appariements complets filtrés (70,17 %)**
    - 365 donnent le bon résultat visuellement (68,48 %),
    - 9 donnent un résultat incorrect (1,69 %). L'algorithme de plus court chemin n'a pas sélectionné le « bon » chemin (choix d'une contre-allée (annexe figure 97) au lieu de la route principale, sens de communication incorrect et non détecté, sélection d'un noeud d'entrée ou de sortie erroné).

- **38 non appariements géométriques**. Les tronçons BD CARTO n'ont pas de tronçons GEOROUTE appariés **(7,13 %)** (tronçon en bordure, très petit tronçon BD CARTO dû au scannage et à la vectorisation (annexe figure 98), tronçon GEOROUTE manquant).

- **121 appariements géométriques où il manque au moins un chemin dans un sens (22,70 %)**
    - 10 cas où il manque un tronçon (tronçon litigieux) dans le chemin correspondant au tronçon BD CARTO (1,88 %), c'est une erreur du processus,
    - 73 cas où il manque des tronçons pour former un chemin (il manque des tronçons GEOROUTE, distances entre tronçons supérieures à 30 m, tronçon BD CARTO en bordure) (13,70%),
    - 21 cas où il manque un noeud de départ ou d'arrivée (appariement des noeuds incohérents) (3,94 %),

- 17 cas où le sens de communication des tronçons est incohérent (3,19 %).

Pour résumer, pour cette zone, le processus d'appariement des tronçons renvoie :
- 68,48 % appariements corrects,
- 1,69 % appariements incorrects (erreurs du processus),
- 27,95 % incohérences dues aux données,
- 1,88 % incohérences dues au processus (erreur du processus).

Pour les incohérences dues aux données, il faut nuancer ce résultat. Effectivement 45 % de ces incohérences sont dues à des tronçons de la BD CARTO qui sortent de l'emprise de GEOROUTE et donc n'engendreront pas de corrections. De plus, la correction des incohérences au niveau des noeuds va résoudre environ un quart des incohérences au niveau des tronçons.

Visuellement, nous avons constaté qu'une correction devra être réalisée dans environ 7% des cas dans GEOROUTE et dans 2% des cas dans la BD CARTO.

#### 5.2.5.4 Conclusion de l'évaluation

Les résultats du processus d'appariement sont donc largement satisfaisants.

Le problème majeur ne provient pas du processus d'appariement mais d'un **mode d'extraction** des données à partir d'un rectangle englobant, différent pour les deux bases :
- pour la BD CARTO, les données livrées sont les instances dont au moins une partie de la géométrie est incluse dans le rectangle,
- pour GEOROUTE, les données livrées sont :
    - les instances dont la géométrie est incluse dans le rectangle,
    - les nouvelles instances issues des données dont la géométrie intersecte le rectangle englobant. La sémantique d'une nouvelle instance est égale à la sémantique de la donnée initiale, et la géométrie est égale à la partie de la géométrie initiale à l'intérieur du rectangle.

Le choix d'un unique mode d'extraction permettrait de diminuer fortement (≈ 40 %) le nombre d'incohérences détectées.

Néanmoins, quelques améliorations sont envisageables. Elles vont maintenant être présentées.

### *5.2.6 Extension du processus*

Ce processus d'appariement pourrait être encore amélioré en utilisant des mécanismes plus puissants ou plus proches de ceux utilisés pour apparier visuellement.

Cinq extensions possibles vont être présentées :
- la prise en compte des tronçons litigieux (5.2.6.1),
- le filtrage par un algorithme de plus proche chemin (5.2.6.2),
- le filtrage de l'appariement des noeuds par l'appariement des tronçons (5.2.6.3),
- l'utilisation de la classe CARREFOUR_COMPLEXE de GEOROUTE (5.2.6.4),
- le paramétrage automatique du rayon de la zone de recherche (5.2.6.5).

#### 5.2.6.1 Prise en compte des tronçons litigieux

Actuellement, les tronçons litigieux de GEOROUTE peuvent uniquement être appariés comme des tronçons composants d'un carrefour complexe, ils ne sont pas retenus lors de l'appariement des tronçons.

L'utilisation d'outils topologiques (plus court chemin, groupe connexe) et des relations topologiques permettrait de résoudre le litige et ainsi d'apparier un tronçon litigieux géométriquement avec un tronçon de la BD CARTO. Par exemple, si les tronçons connexes, appariés au tronçon GEOROUTE litigieux, sont tous appariés avec le même tronçon BD CARTO alors le tronçon GEOROUTE litigieux est apparié avec ce tronçon BD CARTO.

De cette façon, le nombre d'incohérences dues au processus serait diminué. Par contre, le contrôle de cohérence devra vérifier qu'un tronçon litigieux n'est pas apparié avec plusieurs tronçons de la BD CARTO.

#### 5.2.6.2 Filtrage des tronçons de route par un algorithme de plus proche chemin

Pour apparier les tronçons, un algorithme de plus court chemin est utilisé, cet algorithme donne de bons résultats. Cependant dans certains cas, il est mis en défaut. Pour la figure 76 (a), l'algorithme du plus court chemin choisit la ligne droite (figure 76 (b)).

Pour filtrer les tronçons candidats, l'utilisation du **plus proche chemin** (5.1.1.5) prenant en compte le tronçon de la BD CARTO, serait préférable. Ainsi, le chemin sélectionné ne serait plus le chemin le plus court entre les deux extrémités mais le chemin le plus proche du tronçon de la BD CARTO entre les deux extrémités (la figure 76 (c)).

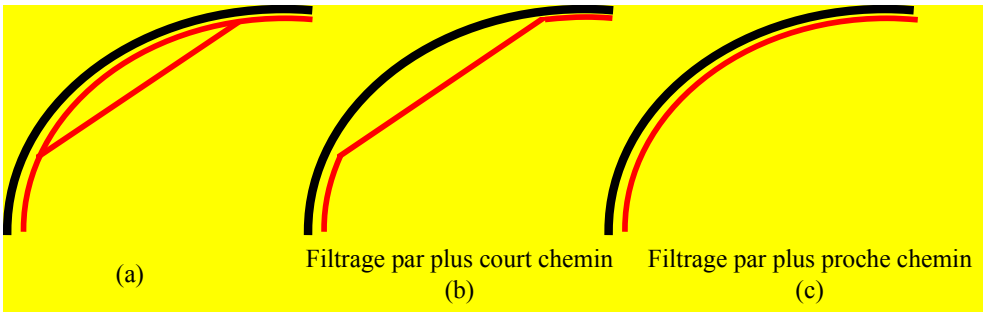



figure 76 : Filtrage par plus court et plus proche chemin

#### 5.2.6.3 Filtrage des appariements de noeuds par l'appariement des tronçons

Le filtrage de l'appariement des noeuds peut aussi être amélioré à l'aide du filtrage de l'appariement des tronçons de routes. En effet, si des points de liaison sont supprimés lors du filtrage de l'appariement des tronçons, ces points peuvent être supprimés pour le filtrage par plus court chemin des noeuds. Les éléments du carrefour ne faisant plus partie des plus courts chemins seront alors retirés de l'appariement. Une fois ces éléments supprimés, l'étape de réduction (5.2.3.4.2.4) pourrait alors être relancée.

Cette extension possible est une reprise de l'étape de filtrage pour des appariements déjà réalisés.

#### 5.2.6.4 Utilisation de la classe CARREFOUR_COMPLEXE de GEOROUTE

Les noeuds de la BD CARTO sont appariés soit avec un noeud, soit avec un carrefour complexe de GEOROUTE. Or, dans GEOROUTE, il existe une classe CARREFOUR_COMPLEXE. Cette classe pourrait être utilisée pour faciliter la construction du carrefour complexe, en utilisant les instances de cette classe dans la zone de recherche comme des groupes candidats. Ces groupes n'auraient pas besoin d'être filtrés par suppression (ils n'ont pas de parasites). Néanmoins, tous les ensembles connexes appariés avec des noeuds

de la BD CARTO ne sont pas tous des instances de la classe CARREFOUR_COMPLEXE (entre autres les petits ronds-points).

#### 5.2.6.5 Paramétrage automatique du rayon de la zone de recherche

Actuellement, le rayon de la zone de recherche est fixé empiriquement en déterminant visuellement le rayon nécessaire pour chaque type de noeud de la BD CARTO. Pour éviter cette recherche empirique, un **rayon variable** pourrait être défini (figure 77). Il est fonction de la distance au noeud le plus proche de la même base, d'un seuil maximum et d'un seuil minimum. Ce paramétrage ajoute cependant une difficulté supplémentaire qui réside dans le choix d'un seuil minimal et d'un seuil maximal. De plus, il ne résout pas les appariements n-m. Cependant, cette solution a été retenue dans [Branly 97] pour apparier les carrefours de la BD TOPO et de GEOROUTE.

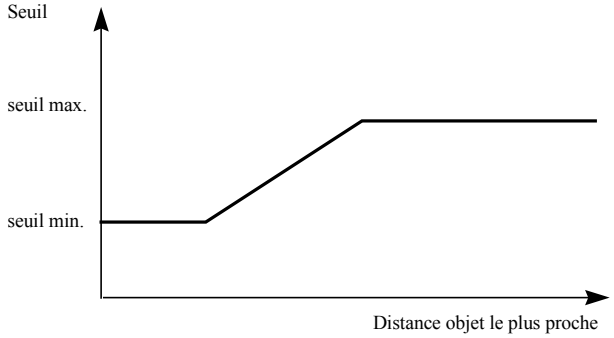



figure 77 : Seuil variable

### *5.2.7 Conclusion sur le prototype d'appariement de données routières*

Pour les données routières de la BD CARTO et de GEOROUTE, un prototype spécifique a été créé. Il permet :

- des appariements complexes de type 1-n,
- la détection des incohérences entre les deux BDG.

Des résultats fiables sont obtenus grâce à la combinaison d'un grand nombre d'informations sémantiques, géométriques et topologiques.

Ce processus d'appariement de BDG routières à différentes échelles peut donc facilement être **généralisé** pour apparier d'autres données de type réseau à différentes échelles. Par exemple, l'appariement de deux **réseaux routiers** au **1 : 1 000 000** et au **1 : 250 000**. Un processus d'appariement pour les réseaux **ferrés** ou **hydrographiques** à différentes échelles est de même facilement dérivable. Par contre, il n'est pas possible d'étendre ce processus pour des données à la même échelle (ce processus ne permettant pas les appariements n-m dus à des spécifications différentes). Un autre processus [Branly 97] respectant le processus générique, a été conçu pour ce cas de figure.

Notre processus a suivi les phases du processus générique d'appariement. Seule la phase de **regroupement** n'a pas été intégrée dans ce prototype. L'utilisation d'outils fournissant des regroupements du même type que les appariements (1-n) désirés a permis d'éviter cette phase. Cependant, le regroupement s'est avéré indispensable pour les processus d'appariement décrit dans [Branly 97] ou [Bucaille 96]. La définition du processus générique d'appariement s'appuyant sur la boîte à outils est donc profitable pour développer plus rapidement d'autres processus d'appariement et réutiliser les outils existants.

## 5.3 Enrichissements par extraction des opérations de généralisation

L'appariement permet de déterminer les objets représentant le même phénomène du monde réel, mais aussi **d'extraire des connaissances multi-représentations** [Devogele et Raynal 95] [Devogele et Raynal 96] qui correspondent aux **différences entre ces représentations**. Celles-ci se trouvent à trois niveaux :

- au niveau **sémantique**, les différences sont dues :
    - aux **disparités entre les schémas** qui peuvent être décrites par des opérations semblables aux opérations d'évolution de schéma décrites dans [Scherrer et al. 93] [Scholl and Tresch 93].
    - aux **conflits de données** qui se détectent facilement par l'inégalité des valeurs entre les objets appariés. Ces inégalités sont corrigées et n'ont donc pas besoin d'être signalées.
- au niveau **topologique**, les différences sont dues :
    - aux **appariements 1-n**, les liens topologiques entre les n objets ne sont pas fournis par l'autre représentation (4.4.2),
    - à des **incohérences** qui sont décelées en comparant les instances des relations topologiques entre les objets appariés. Ces incohérences sont corrigées et n'ont donc pas besoin d'être signalées.
- au niveau **géométrique**, la récupération des différences est plus difficile et plus enrichissante, ces différences sont dues :
    - à des incohérences,
    - à des variations entre les abstractions des différentes représentations des phénomènes du monde réel. Pour qualifier ces différences, nous utiliserons les **opérations de généralisation**[19], définies dans [Shea et McMaster 91] [McMaster and Shea 92] [Affholder 95], [Le Men 96] ou [Peng, et al. 96], comme langage de description de la différence entre les objets représentant le même phénomène du monde réel.

Dans cette partie, nous allons uniquement décrire les différences au niveau **géométrique** et **comment extraire les opérations**. Les opérations pour le prototype routier seront présentées dans la section 5.3.2. Les autres seront abordées dans la section 5.3.3. Pour chaque opération, les critères permettant de la déduire seront fournis. De plus, le terme anglais usuellement consacré, sera donné entre parenthèses. Mais auparavant, les utilisations possibles de ces opérations vont être exposées en 5.3.1.

### *5.3.1 Apports des opérations de généralisation*

Les opérations de généralisation sont précieuses pour propager les mises à jour, pour l'aide à la généralisation de BDG et pour contrôler la cohérence de deux bases.

Pour **propager les mises à jour** [Kilpeläinen 95] [Kemppainen 94] des représentations les plus fines vers les représentations les moins détaillées, les opérations de généralisation peuvent être mises à profit, dans une certaine mesure. Effectivement, si les correspondances portent assez d'informations (séquence des opérations précises, paramètres…), la mise à jour pour une représentation moins détaillée, obtenue par conversion à partir de la mise à jour du niveau le plus détaillé, peut être déclenchée. Cependant, il est clair que pour des mises à jour géographiques complexes ou des mises à jour impliquant plusieurs éléments, cette automatisation sera difficile (2.1.6). Pour ces cas complexes, la séquence d'opérations de

[19] opérations qui consistent à modifier les données afin d'obtenir une représentation plus simple et plus abstraite

généralisation choisie risque d'être différente. Néanmoins, les opérations de généralisation sont d'une grande aide même pour un processus de mise à jour interactif.

Pour la **généralisation de BDG**, ces opérations inférées sont intéressantes car cet apprentissage peut conduire à des connaissances réutilisables par analogie. En effet, les opérations détectées peuvent servir pour déterminer les opérations de généralisation candidates, en fonction des objets à généraliser et du contexte. L'analyse des séquences d'opérations ainsi détectés peut aussi contribuer à la progression des stratégies de généralisation automatique. La base de données multi-représentations servira alors de **plate-forme** de recherche pour la généralisation automatique.

Pour le **contrôle de cohérence**, les modifications stockées dans les opérations (suppression, déplacement, amalgamation) vont pouvoir être employées pour vérifier les spécifications de saisie. Ainsi, une vérification des contraintes (déplacement maximum, contrôle sur l'amalgamation d'objets de types différents,…) définies sur la BDG la moins détaillée sera réalisable. La BD la plus détaillée servira alors de référentiel.

Plus simplement, les opérations détectées peuvent aider l'utilisateur à comprendre la signification des différences entre les données. Par exemple, pour des correspondances entre un ensemble d'habitations et une zone d'habitation, nous allons savoir qu'une amalgamation est l'opération « équivalente » aux différences constatées, de même entre deux tronçons de route, nous avons besoin de savoir que la généralisation équivalente est par exemple, un lissage avec un déplacement.

### *5.3.2 Opérations de généralisation inférées pour le prototype routier*

Dans cette section, les opérations pouvant être déduites du prototype routier seront exposées. Pour une même correspondance, plusieurs opérations ou séquences peuvent être déduites.

**Suppression (*deletion*)**

L'opération la plus simple est la suppression. Elle consiste à supprimer la représentation du phénomène du monde réel dans la BDG la moins détaillée. Une opération de suppression se déduit par l'absence de données correspondantes. Elle peut être appliquée à n'importe quel type d'objet (ponctuel, linéaire, surfacique).

**Filtrage (*simplification*)**

Le filtrage consiste à supprimer des points intermédiaires dans une polyligne. Une opération de filtrage est détectée lorsque la ligne de l'objet en correspondance a un nombre de points intermédiaires inférieur à la ligne du premier objet. Elle peut aussi être appliquée à la frontière d'une surface.

**Lissage (*smoothing*)**

Cette opération transforme une polyligne en une autre, qui se trouve décalée localement vers le centre de la courbure. Elle se caractérise par une ligne moins anguleuse. Cette opération a pour propriété que la somme des angles de l'objet généralisé est inférieure à la somme des angles de l'objet apparié.

**Déplacement (*displacement*)**

L'opération de déplacement correspond à une translation d'un objet selon un vecteur.

Pour un **objet ponctuel**, le déplacement est paramètrés par le vecteur de translation. Un déplacement est déduit quand les points des objets en correspondance n'ont pas les mêmes coordonnées.

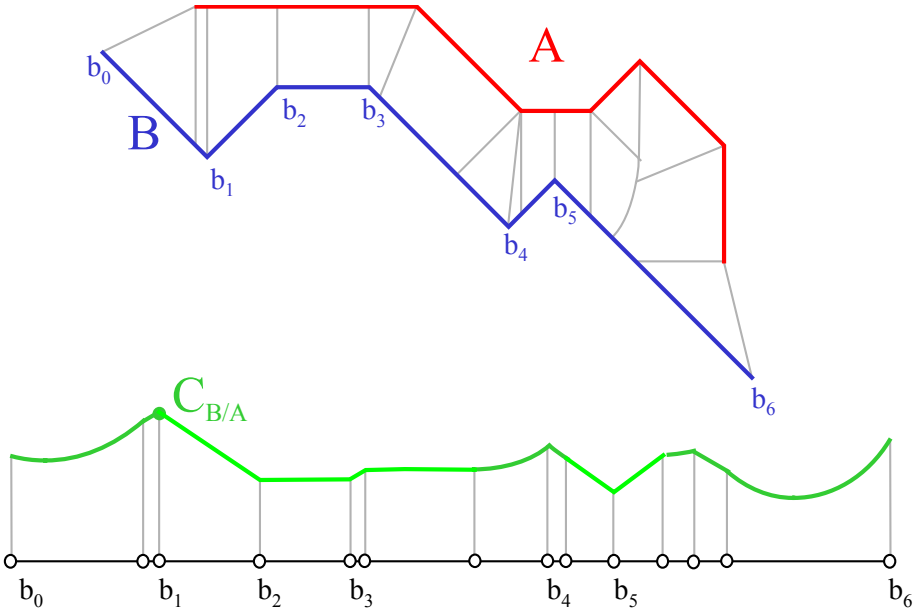



figure 78 : Fonction de distance de l'objet B vers l'objet A

Pour **un objet linéaire**, le déplacement est théoriquement caractérisé par un vecteur de déplacement qui est affecté à tous les points de la ligne. Les lignes étant issues de sources différentes, il n'existe pas de vecteur de translation constant pour l'ensemble des points. Un paramétrage plus fin doit être défini pour caractériser le déplacement. Dans cet objectif, la fonction de mesure d'une des deux composantes de Hausdorff obtenue à l'aide du diagramme de Voronoï [Alt et al. 92] est utilisée. Cette fonction donne pour chaque point de la ligne B, la distance minimale à la ligne A. Pour la figure 78, le déplacement de B vers A peut être caractérisé par la fonction distance de B vers A à l'aide du diagramme de Voronoï (dont une partie est dessinée en gris entre A et B).

Si, pour une ligne, les variations de la fonction distance sont importantes, une opération de caricature peut aussi être extraite.

**Caricature (*exaggeration*)**

La caricature est une opération qui augmente la taille d'une partie de l'objet, en privilégiant une direction (figure 79 : Exemple de caricature). Elle est utilisée pour conserver un détail de la ligne trop petit pour être représenté à la résolution finale et trop important pour être supprimé.

figure 79 : Exemple de caricature

Pour le linéaire, cette opération se caractérise par une distance moyenne faible et une distance maximale prenant une valeur éloignée de la distance moyenne. Une caricature peut être combinée avec un déplacement. Dans ce cas, la distance moyenne est importante et la distance maximale est très différente de la distance moyenne. Cette opération peut aussi être appliquée à des objets surfaciques.

**Fusion linéaire**

Toutes les opérations présentées jusqu'à maintenant ne traitaient qu'un objet, les opérations utilisant plusieurs objets vont maintenant être abordées. La fusion linéaire est la plus couramment utilisée. Cette opération transforme une suite de tronçons formant un chemin en un tronçon unique. Cette opération est facile à détecter, elle se caractérise par une correspondance 1 tronçon et n tronçons formant un chemin.

**Composition**

Une composition est une opération sur les objets qui consiste à transformer un ensemble d'objets formant un objet complexe en un objet simple avec une géométrie qui est issue de la géométrie des objets simples. Nous passons donc d'une représentation détaillée à une « macro » représentation qui regroupe des objets à des niveaux moins détaillés.

Par exemple, un ensemble de tronçons et de noeuds formant un carrefour complexe peut être généralisé en un noeud routier.

L'opération de composition se caractérise par une correspondance 1-n et des objets initiaux et finaux de types différents. Le changement de dimension d'une partie ou de la totalité des instances est très probable. Cette opération est possible pour des objets ponctuels (ensemble de ruines ponctuelles vers zones en ruine), linéaires (ensemble de voies de chemin de fer linéaire vers zone de triage) ou surfaciques (ensemble de maisons vers zone d'habitation).

**Amalgamation linéaire ou unification (*merge*)**

L'amalgamation linéaire ou unification est une composition, elle transforme des tronçons « parallèles » en un tronçon. [Shea et McMaster 91] donne l'exemple de grandes routes à voies séparées, normalement représentées par une ou plusieurs lignes adjacentes, avec une distance de séparation entre elles. Par réduction d'échelle, ces deux lignes sont fusionnées en une seule située approximativement à mi-chemin entre les deux lignes initiales. L'amalgamation linéaire peut être déduite quand des tronçons non connexes et « parallèles » sont appariés à un tronçon. Plus généralement, quand un ensemble de tronçons non connexes et« parallèles » sont appariés avec un ensemble de tronçons formant un chemin.

### *5.3.3 Autres opérations inférées possibles*

D'autres opérations ne concernant pas les réseaux routiers peuvent aussi être détectées par un processus d'appariement. Les opérations sur les objets de type réseau (hydrographique, ferré,…) sont sensiblement les mêmes que celles décrites pour le prototype.

Par contre, pour les autres thèmes (occupation du sol [Le Men 96],…) la détection d'autres opérations est envisageable. Nous allons décrire ces extractions.

**Exagération surfacique**

L'exagération surfacique consiste à augmenter la taille de l'objet surfacique, elle est employée pour des objets trop petits pour être représentés à la résolution donnée, et trop importants pour être supprimés. Elle agit de façon égale dans toutes les directions sur l'ensemble de l'objet, ce qui la différencie de la caricature. Une opération d'exagération surfacique est déduite, si la surface de l'objet apparié est plus importante que la surface de l'objet initial et si la fonction distance (figure 78) entre les contours des objets appariés a une variation faible.

**Fusion surfacique ou agglomération (*merge*)**

L'opération de fusion (encore appelée agglomération) consiste à regrouper en un objet de même nature un ensemble connexe d'objets (figure 80). Cette opération est souvent la conséquence d'une opération sur le schéma (suppression d'attributs,…). Celle-ci peut être déduite si un ensemble d'objets surfaciques connexes est apparié à un seul objet surfacique.

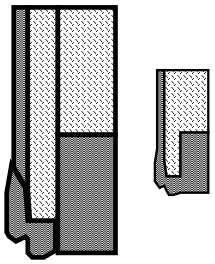


figure 80: Exemple de fusion de parcelles

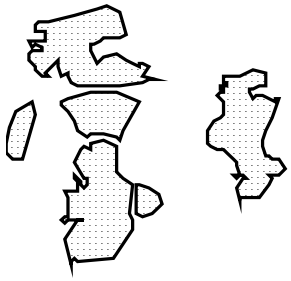


figure 81: Exemple d'amalgamation de "petites" forêts

**Amalgamation surfacique ou jonction (combine, amalgamation)**

L'opération d'amalgamation (encore appelée jonction) consiste à regrouper en un objet surfacique de même nature un ensemble non connexe d'objets (figure 81). Cette opération regroupe un ensemble d'objets dont l'inter-distance est trop petite pour la **résolution**. Elle se caractérise par une correspondance 1-n et une non connexité des n objets en relation. Cette opération peut être inférée si un ensemble d'objets surfaciques non connexes est apparié à un objet surfacique.

**Changement de dimension (*collapse*)**

L'opération de **changement de dimension** modifie la dimension de la géométrie de l'objet. Par exemple, une rivière surfacique devient une rivière linéaire. Cette opération est facile à retrouver, il suffit de comparer la dimension des géométries associées aux objets géographiques appariés.

**Structuration (*typification*)**

La structuration est une opération de simplification spatiale qui consiste à conserver l'expressivité globale aux dépens de la cardinalité et de l'expressivité élémentaire (figure 39 page 3). Elle est utilisée quand la diminution de la résolution ne permet plus de garder tous les éléments d'un groupe d'objets de même type et de mêmes valeurs sémantiques. En généralisant, l'ensemble d'objets est alors réduit à un ensemble plus petit. Ce dernier n'est pas un sous-ensemble. Un des objets structurants en lui-même n'a aucune signification ; par contre, l'ensemble, par sa forme, l'espacement entre les objets et la position de ceux-ci, permet de traduire l'aspect général des objets de départ. Cette représentation a perdu son caractère cardinal et l'expressivité élémentaire. Une structuration peut être déduite quand il existe une correspondance entre deux ensembles d'objets qui représentent le même phénomène, sans qu'il existe, dans ces ensembles, des correspondances entre un objet structurant et un objet structuré.

**Désagrégation**

La **désagrégation** est une autre opération utilisée en occupation du sol. Elle consiste à découper l'objet surfacique en plusieurs objets pour traduire au mieux la forme de l'objet (figure 82). Cette opération est utilisée quand une section de la géométrie de l'objet est inférieure au seuil fixé. Une désagrégation peut être inférée quand il existe une correspondance 1-n et une non connexité des n objets appariés.

figure 82: Exemple de désagrégation

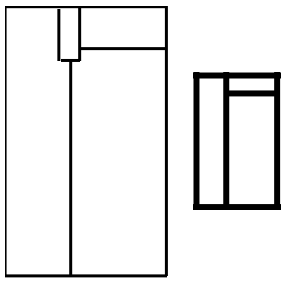


figure 83: Exemple de destruction/partage

**Destruction/partage**

L'opération de **destruction/partage** [Le Men 96] (figure 83) est proche de celle de désagrégation. Mais elle est contrainte par une partition totale de l'espace. Elle est utilisée pour répartir la surface laissée par un objet retranché entre ses voisins. Cette opération est couramment utilisée en occupation du sol. Elle se déduit par une correspondance n-m, les n et m instances surfaciques sont connexes et ont une emprise similaire.

### *5.3.4 Conclusion sur l'enrichissement par extraction des opérations de généralisation*

Cette partie a présenté un ensemble d'opérations pouvant être inférées par le processus d'appariement. Ces opérations sont précieuses pour améliorer le processus de généralisation, la propagation des mises à jour et le contrôle de cohérence. D'autres opérations plus complexes peuvent être aussi détectées comme la schématisation de virages ou l'accordéon [Plazanet 96]. Le nombre d'opérations extraites est fonction de l'utilisation de cet ensemble, des opérations que l'utilisateur désire inférer. Quoi qu'il en soit, pour être utilisées efficacement, ces opérations doivent être caractérisées par des **paramètres**. Certains de ces paramètres, comme la distance moyenne, ont déjà été présentés, d'autres comme la détection de la direction d'une caricature, devront être récupérés par d'autres outils.

## 5.4 Conclusion sur l'appariement

L'intégration des BDG doit, la plupart du temps, inclure un processus d'appariement. Ce chapitre a proposé un **processus générique** qui définit les phases communes au processus d'appariement (enrichissement, sélection, mesure, filtrage ou prolongation, regroupement, analyse) et leur enchaînement. Ce processus s'appuie sur une boîte à outils d'appariement. Ces outils permettent de comparer la sémantique, la géométrie et les liens topologiques des objets candidats à l'appariement. Afin d'illustrer ce processus générique, le prototype utilisé pour apparier les données routières de la BD CARTO et de GEOROUTE a été présenté. Le processus d'appariement permet aussi d'extraire des connaissances multi-représentations décrivant les différences entre les objets appariés à l'aide d'opérations de généralisation. Ces connaissances en sont encore à l'état d'ébauche.

Les processus d'appariement sont complexes, ils confrontent un grand nombre d'informations et enchaînent plusieurs d'outils. La conception d'un processus d'appariement est donc une tâche fastidieuse. Néanmoins, la définition d'un processus générique facilite largement cette tâche. Cependant, la conception de ceux-ci peut encore être simplifiée en assistant l'utilisateur dans :

- le choix des outils d'appariement,
- le paramètrage de ces outils,
- le choix des filtres,
- l'enchaînement des outils.

Pour ce faire, des techniques d'apprentissage par induction [Gascuel 87] pourraient aussi être utiles.

# 6. Conclusion

Pour conclure, nous allons résumer les contributions majeures de cette thèse (6.1) et dégager les perspectives principales (6.2).

## 6.1 Contribution de la thèse

Trois apports essentiels découlent de cette thèse. Tout d'abord, une taxonomie des conflits d'intégration pour les BDG a été établie (6.1.1). Puis, grâce à cette taxonomie, un processus d'intégration des BDG vectorielles (6.1.2) a été défini. Celui-ci fait appel à un processus d'appariement (6.1.3), ce qui constitue le troisième apport de cette thèse.

Le processus d'intégration / appariement défini dans cette thèse, a permis d'obtenir une BDG multi-représentations à partir des trois principales bases de données de l'IGN (BD TOPO, BD CARTO et GEOROUTE) pour le thème routier dans la région de Lagny. Cette BDG multi-représentation répond aux besoins des applications multi-représentations.

### *6.1.1 Taxonomie des conflits d'intégration pour les BDG*

Plusieurs taxonomies ont déjà été réalisées pour les BD classiques. Mais, jusqu'à présent, aucune n'avait pris en compte les spécificités de l'information géographique. Cette taxonomie [Spaccapietra et al. 96] [Parent et al. 96] a ainsi permis de recenser les problèmes à résoudre pour intégrer les BDG. Elle étend les taxonomies réalisées pour les BD classiques et inclut les problèmes d'intégration des BDG décrit dans les articles précédents ([Gouvernement du Québec 92] [Rigaux et Scholl 95] [Shepherd 92] [Laurini 96]). Cette taxonomie a fait apparaître un grand nombre de conflits supplémentaires liés à la nature de l'information géographique, il en résulte six classes :

1. les **conflits de sources de données** qui apparaissent quand les types de sources de données employées ou les caractéristiques de celles-ci sont différentes,
2. les **conflits d'hétérogénéité** qui portent sur les critères globaux à définir pour chaque BDG (modèles de données, positionnement des éléments, modélisation de l'altitude, mode de représentation, méta-données liées aux géométries et relations topologiques),
3. les **conflits de définition des classes** qui regroupent les problèmes liés à la classification des éléments des BDG, à leur spécification et au découpage des phénomènes du monde réel en objet,
4. les **conflits de structures** qui surviennent lorsque les éléments en correspondance sont décrits par des concepts différents (classe, attribut, relation) ou lorsque qu'une information gérée par la BDG correspond à une information qui doit être déduite,
5. les **conflits de description** sémantiques et géométriques qui résultent des différences entre les propriétés (attributs, méthodes) des classes en correspondance,
6. les **conflits de données** qui surviennent lorsque les objets en correspondance ont des valeurs différentes pour des attributs en correspondance.

### *6.1.2 Définition d'un processus d'intégration de BDG*

Plusieurs processus d'intégration de BDG avaient déjà été proposés, mais ils étaient incomplets ou se limitaient à certains cas particuliers. La taxonomie des conflits d'intégration a permis de définir pour les BDG vecteurs, un processus générique d'intégration [Devogele et al. 97]. La complexité des conflits rencontrés nous a poussé à choisir un processus déclaratif en trois phases afin de le décomposer au maximum en sous-tâches. Pour cela, nous avons

repris chacune des 3 phases définies dans le processus d'intégration de [Parent et Spaccapietra 96] et nous les avons adaptées afin de tenir compte des spécificités des BDG. Il se décompose comme suit :

- pré-intégration,
- déclaration des correspondances,
- intégration.

#### 6.1.2.1 Définition du processus de pré-intégration pour les BDG

La pré-intégration inclut toutes les activités préliminaires qui ont pour objectif de faire converger les descriptions initiales. Elle consiste à réarranger les schémas en entrée pour les rendre plus homogènes sur le plan sémantique et syntaxique pour parvenir au même niveau de compréhension des données.

La première tâche consiste à choisir un **modèle de données** pour la BD intégrée. Le modèle **orienté objet** a été choisi afin de bénéficier de sa capacité à modéliser des phénomènes complexes. De même, un **système de référence** doit être retenu durant cette phase.

La deuxième tâche a pour but d'enrichir les BDG à l'aide de méta-données, de mécanismes de traduction et de nouvelles données. Ces données sont les informations implicites qui ont été matérialisées pour mettre les BD en conformité.

Enfin, dans la dernière tâche, les **règles de normalisation** spécifiques aux BDG qui ont été définies (suppression des restrictions, niveau de décomposition homogène, suppression des classes d'objets géométriques), sont appliquées à l'ensemble des BDG à intégrer pour les rendre plus homogènes.

#### 6.1.2.2 Extension du langage de déclaration des correspondances

L'objectif de la déclaration de correspondance est d'identifier et de fournir toutes les correspondances entre les schémas de données à un niveau sémantique et entre leurs instances. Elle s'appuie sur **la déclaration d'assertions de correspondance inter-schémas (ACI)** qui mentionnent les éléments en correspondance.

Pour intégrer les BDG, la syntaxe des **déclarations de correspondance** a été étendue, afin de permettre la déclaration des conflits d'intégration propres aux BDG. Ces extensions sont novatrices, car elles autorisent la déclaration de conflits de spécifications (4.3.3.1), de granularité (4.3.5.1) et proposent une syntaxe relativement simple pour les conflits de classification (4.3.1.1) et de fragmentation (4.3.2.1).

Afin de les déclarer, les assertions incluent des attributs virtuels, des classes virtuelles et autorisent l'expression de critères de sélection. Une déclaration au niveau des types des attributs a aussi été définie pour permettre une intégration optimale des attributs en correspondance faible (4.3.4.1.3).

#### 6.1.2.3 Extension des techniques d'intégration

L'intégration proprement dite traite toutes les ACI déclarées dans la phase précédente. Elle résout les conflits décrits dans les déclarations, fournit le schéma de la base intégrée et produit les règles de traduction des schémas initiaux vers le schéma intégré, et les règles réciproques.

Cependant, avant d'intégrer les BDG, une stratégie d'intégration doit être choisie en fonction de l'objectif de l'intégration. Nous avons ainsi défini deux stratégies d'intégration (mono-représentation et multi-représentation). La stratégie **mono-représentation** (4.3.1.2.1) produit une représentation unique des phénomènes du monde réel en retenant l'information la

plus précise sans redondance. La stratégie multi-représentation (4.3.1.2.2) par contre, conserve les représentations à différents niveaux de détail et les relie entre elles.

Pour chaque conflit et chaque stratégie, nous avons déterminé une technique d'intégration. Ces techniques permettent d'intégrer les classes, les attributs, les géométries et les relations en correspondance. Elles s'appuient sur les techniques d'intégration de schéma classiques et de fusion des objets géographiques. Celles-ci ont été améliorées pour maintenir l'ensemble de l'information initiale.

### *6.1.3 Définition d'un processus d'appariement*

Le processus d'intégration de BDG doit le plus souvent inclure un processus d'appariement qui établit des correspondances fiables entre les objets représentant le même phénomène du monde réel, à l'aide un grand nombre d'informations (sémantique, géométrique et topologique) et de plusieurs outils.

Dans le domaine de l'appariement, trois apports sont à noter :

- la définition **d'une boîte à outils** regroupant les outils nécessaires au processus d'appariement. Ces outils ont été organisés en cinq groupes (outils sémantiques, outils topologiques, outils géométriques de type distance, outils géométriques de type forme, outils mixtes) en fonction de la caractéristique qu'ils mesurent.
- la définition d'un **processus générique** à partir de l'analyse des différents processus d'appariement réalisés au COGIT ([Bucaille 97] [Branly 97] [Devogele et al. 96 a]). Il emploie la boîte à outils d'appariement précédemment citée. Le but de ce processus générique est de définir les **phases communes** à un processus d'appariement quelles que soient les données à apparier et de déterminer leur **enchaînement**.

Le processus générique a été illustré dans cette thèse par le prototype développé pour apparier les données routières de la BD CARTO et GEOROUTE sous $GéO_2$, avec des résultats concluants.

- la **qualification des appariements** obtenus **par extraction d'opération de généralisation** pour aider les processus de contrôle de cohérence, de propagation des mises à jour ou d'apprentissage (par analogie) pour la généralisation. Cette qualification est encore à l'état d'ébauche néanmoins, les opérations peuvent déjà être inférées à l'aide des mesures réalisées lors du processus d'appariement.

## 6.2 Perspectives

Les perspectives de recherches complémentaires ou ultérieures sont nombreuses. Dans la section 6.2.1, les améliorations qui peuvent être apportées seront développées, puis la section 6.2.2 montrera comment tirer profit des BDG multi-représentations issues du processus d'intégration / appariement réalisé, pour les nouvelles applications utilisant cette spécificité. Finalement, dans la section 6.2.3, les extensions nécessaires pour autoriser l'intégration de BDG réparties sur un réseau (BD fédérées) seront présentées.

### *6.2.1 Extension et amélioration du processus d'intégration / appariement*

Certaines étapes ou certains points du processus d'intégration / appariement restent délicats et peuvent donc encore être améliorés ou étendus. Nous aborderons la fusion des données incompatibles, l'amélioration du processus d'appariement, et l'ajout des éléments manquants pour assurer la complétude de ce processus.

#### 6.2.1.1 Fusion des données incompatibles

La fusion des données géographiques fait appel à des mécanismes de fusion. Or, ceux réalisés dans le cadre de cette thèse ne sont pas satisfaisants, pour résoudre les conflits de fragmentation ou les conflits de critère de décomposition.

La fusion d'objets en **conflit de fragmentation n-m** (figure 49 page 3) est conditionnée par la conception d'un mécanisme de **désagrégation** des valeurs des attributs [Flowerdew et Openshaw 87] [Weber 94]. Il serait donc nécessaire de poursuivre les recherches dans ce domaine.

De même, pour intégrer les objets en **conflit de critère de décomposition** (figure 28 page 3), de nouveaux mécanismes doivent donc être conçus. Ils doivent tenir compte du contexte (géométries des objets voisins, …) et permettre de déduire une géométrie, à partir des géométries initiales. Il est clair que les progrès en ce domaine dépendront des progrès relatifs à la modélisation et à la maîtrise de la qualité géométrique.

#### 6.2.1.2 Processus d'appariement « intelligent »

Des améliorations concernant le processus d'appariement peuvent aussi être réalisées. La définition d'un processus générique est un premier pas fondamental pour l'appariement. Cependant, trois améliorations sont possibles :

- en **aidant** l'utilisateur à **choisir les outils, leurs paramètres et leur enchaînement**. Ainsi, un ordre de sélection pourrait être déterminé par le processus en fonction des AIC. De même, des techniques d'apprentissage par induction [Gascuel 87] pourront aussi être utilisées pour aider l'utilisateur dans son choix des outils et des valeurs des paramètres. Un petit nombre d'appariements serait alors réalisé manuellement, puis le processus déterminerait les outils et les paramètres pour réaliser ces appariements.
- en **rendant le processus d'intégration moins déterministe**. Actuellement, le mécanisme d'appariement utilise une logique classique des prédicats du premier ordre (oui, non) et combine un certain nombre de prédicats pour accepter ou refuser l'appariement. Or, visuellement, nous pouvons accepter des appariements ayant un critère d'appariement non vérifié si les autres critères d'appariement sont vérifiés. L'introduction d'une **logique floue** [Bouchon-Meunier 94] permettrait de disposer d'un mécanisme d'appariement plus proche de notre appariement visuel (ajout de la notion d'ambiguïté sur l'expression des connaissances et des implications, possibilité de revenir sur un résultat, …).
- en **renvoyant des opérations de généralisation précises** caractérisées par des paramètres. Certains de ces paramètres comme la distance moyenne sont déjà calculables mais d'autres comme la direction de la caricature devront être récupérés par des outils qui restent à définir. De plus, d'autres opérations plus complexes peuvent être aussi détectées comme les opérations de schématisation de virages [Plazanet 96]. Cette direction de travail reste donc largement à explorer et est dépendante des progrès en matière de généralisation automatique.

#### 6.2.1.3 Vers un processus d'intégration de BDG global et automatique

Pour obtenir un processus d'intégration de BDG global et automatique, l'intégration des relations des BDG, des BDG raster et des BDG 3D doit être approfondie. De même, un mécanisme de traduction automatique des schémas initiaux vers le schéma intégré doit être conçu.

Dans cette thèse, **l'intégration des relations** n'a été qu'évoquée. Une étude plus approfondie est nécessaire pour traiter avec précision la déclaration des assertions de correspondance entre les relations, et définir des techniques d'intégration appropriées. Actuellement, hormis la technique de définition de relation virtuelle, l'intégration des relations dans les BDG est limitée à l'intégration des relations dans les BD classiques.

D'une manière plus générale, le processus d'intégration / appariement doit être généralisé aux BDG raster ou 3D. Pour généraliser l'intégration / appariement aux **BDG raster**, des techniques de fusion de géométries raster doivent être définies ; des outils d'appariement spécifiques au raster doivent être intégrés. De même, une extension aux **BDG** gérant la **troisième dimension** (l'altitude) est envisageable. Dans cet objectif, des outils d'appariement de volumes devront être développés, et des techniques d'intégration des conflits de la gestion de la troisième dimension devront être proposées.

Finalement, pour automatiser la troisième phase du processus d'intégration, un **mécanisme de traduction automatique** des schémas initiaux vers le schéma intégré et vice versa, doit être développé. Il n'a pas été réalisé, car il dépend fortement du modèle des BDG utilisés. Actuellement, plusieurs organismes (ISO, CEN, OpenGIS) travaillent à l'élaboration d'un modèle standard et de passerelles vers ce modèle commun. Il sera alors plus facile et plus fructueux de déterminer des mécanismes de traduction automatique des schémas, afin de rendre les modèles **interopérables**.

### *6.2.2 Perspectives pour les nouvelles applications multi-représentations*

Cette thèse a aussi permis de recenser les applications (2.1) qui peuvent bénéficier de la présence de plusieurs représentations, une fois les BDG intégrées. Les principaux outils pour gérer et manipuler les BD multi-représentations vont être présentés pour quelques applications.

#### 6.2.2.1 Perspectives pour la cartographie électronique multi-représentation

Actuellement, les SIG du commerce disposent de quelques outils de visualisation multi-représentation (vues pré-définies, choix de la représentation, symbolisation en fonction de l'échelle). Des techniques de zoom intelligent [Frank et Timpf 94], [Bederson et Hollan 94], ont aussi été implantées dans le cadre de projet de recherche. Elles permettent de changer de représentation lors du changement d'échelle graphique. Les BD multi-représentations permettent de concevoir deux nouvelles applications. la conservation et la propagation des sélections.

La **conservation des sélections** [Timpf et Devogele 97] consiste, lors d'un changement de représentation, à transférer les sélections réalisées par l'utilisateur pour la nouvelle représentation. Par exemple, sur la figure 84, l'utilisateur a sélectionné les objets (en épais), puis il décide de changer de représentation afin de disposer de plus de détails. Les objets de la nouvelle représentation correspondant aux objets sélectionnés dans l'ancienne, doivent alors être sélectionnés.

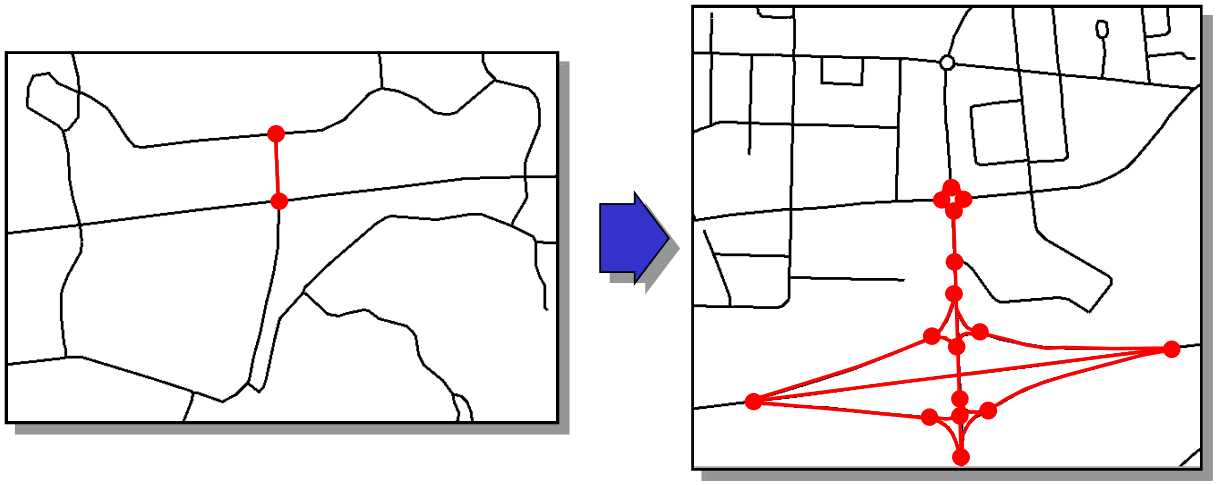


figure 84 : Exemple de conservation des sélections

La **propagation des sélections** [Timpf et Devogele 97] est une opération analogue, elle est employée dans le cadre du multi-fenêtrages. Elle est utilisée pour reporter les sélections réalisées dans les fenêtres actives, aux représentations des fenêtres non actives. Ainsi, les noeuds et les tronçons des représentations des fenêtres non actives correspondants à un noeud sélectionné dans la fenêtre active, seront aussi sélectionnés.

#### 6.2.2.2 Perspectives pour la navigation multi-représentation

L'intégration / appariement permet aussi d'étendre les algorithmes classiques de la théorie des graphes aux représentations multi-représentations. Pour autoriser ces extensions, nécessaires la navigation multi-représentation, la méthode classique « **suivant** » a été surchargée. Normalement**,** cette méthode de la classe noeud renvoie l'ensemble des noeuds reliés à l'instance par un tronçon associé et la longueur du tronçon. La surcharge a consisté à renvoyer des noeuds de différents graphes à l'aide des correspondances entre les noeuds et les tronçons.

Ainsi, pour les graphes de la figure 85, La méthode suivant pour le noeud 'a', va renvoyer les noeuds 'b', 'c', 'd' et les noeuds '5', '6', '7', '8' qui sont les noeuds reliés aux noeuds correspondant à 'a'. Chaque noeud sera couplé avec sa « distance » à 'a', par exemple, pour 'b' la distance est la longueur du tronçon (a, b), pour '5', la distance est la longueur du tronçon reliant le noeud sortant '1' et '5'.

La méthode « suivant » peut être définie sur plus de deux graphes hiérarchiquement ordonnés (par exemple, un graphe entre les villes, un graphe représentant le réseau principal et un graphe représentant l'ensemble du réseau) en utilisant les liens de correspondance entre les noeuds et les tronçons du graphe i et les noeuds et les tronçons du graphe i+1.

La méthode « suivant » définit donc des relations d'un noeud vers l'ensemble des noeuds des différents graphes. Ces relations sont similaires à celles rencontrées entre les noeuds d'un graphe planaire métrique orienté unique. Les algorithmes (plus court chemin [Zhan 96], voyageur de commerce, …) définis sur ce type de graphe peuvent donc être appliqués.

De même, les **collages cognitives** [Claramunt et Mainguenaud 96] permettant de relier des graphes de différents niveaux, sont aussi possibles.


figure 85 : Exemple de liens de correspondance entre les « noeuds »

L'utilisation de plusieurs graphes pour la navigation est avantageuse par rapport à l'utilisation d'un graphe détaillé unique. En effet, elle améliore les performances des algorithmes de navigation sur des « grandes » distances et elle fournit une description des itinéraires plus proche des descriptions naturelles. Cependant, les algorithmes classiques doivent être encore optimisés pour exploiter au mieux l'ensemble des graphes. Par exemple, le choix du graphe doit tenir compte de la distance au point d'arrivée au point de départ, et de la densité du réseau pour définir le chemin entre deux points.

#### 6.2.2.3 Perspectives pour le contrôle de cohérence

Actuellement, les producteurs de BDG sont soucieux de contrôler la qualité de leurs bases et de fournir des indicateurs de qualité. La confiance qu'un utilisateur peut accorder au résultat d'une requête est fonction de ces indicateurs. Le regroupement des différentes représentations dans une seule base intégrée va permettre de répondre à certains de ces besoins.

Premièrement, la BD intégrée va favoriser le **contrôle qualité,** les erreurs contenues vont pouvoir être détectées en comparant les valeurs des instances provenant des différentes bases, puis corrigées.

Deuxièmement, pour les représentations les moins détaillées, les appariements vont permettre de **vérifier la conformité aux spécifications de saisie**. Par exemple, pour la BD1, la spécification de saisie suivante sur les impasses, a été définie : une impasse doit être saisie si sa longueur est supérieure à 100 mètres. Si une impasse de plus de 100 mètres de la BD2 n'est pas appariée avec les instances de la BD1, cette spécification n'est pas respectée.

Des **indicateurs de qualité** vont aussi être fournis pour les représentations les moins détaillées, en prenant la représentation précise comme référence. Par exemple, si les objets d'une zone de la BD TOPO ont été appariés avec les objets de la zone équivalente de la BD CARTO, l'erreur moyenne quadratique des objets linéaires de la BD CARTO peut être réalisée [Bonin 95] (dans cet article un appariement manuel est réalisé avant de contrôler la précision géométrique).

Ainsi, pour le linéaire, la moyenne des distances moyennes va qualifier l'erreur moyenne et le maximum des distances de Fréchet va fournir l'erreur maximale. Par la suite, la concomitance des différentes représentations va faciliter le **maintien de la cohérence** (assurance qualité).

#### 6.2.2.4 Perspectives pour la dérivation de BD hétérogènes

A partir d'une BD multi-représentation constituée à l'aide d'un processus d'intégration / appariement, il est possible de dériver des BD mono-représentations. Pour cela, il suffit pour tous les phénomènes du monde réel ayant plusieurs représentations, de sélectionner une seule représentation. Cependant, une sélection sans contraintes risque de provoquer des incohérences (superposition, topologie défectueuse, …). Il faut donc définir des **contraintes sur les sélections**. Ainsi, dans la figure 86, une BD hétérogène peut être établie en posant les trois règles suivantes :

- La frontière entre les zones ayant des représentations différentes est définie par des tronçons représentés dans les deux représentations.
- Pour les noeuds routiers et les tronçons sur cette frontière la représentation détaillée est sélectionnée.
- Pour raccorder géométriquement les deux graphes, dans la zone peu détaillée, on relie les segments des tronçons ayant comme extrémité un noeud sur cette frontière. Pour GEOROUTE, un raccordement similaire moins automatique a déjà été développé pour relier les données propres à cette base, aux données venant de la BD CARTO [Trevisan 95].

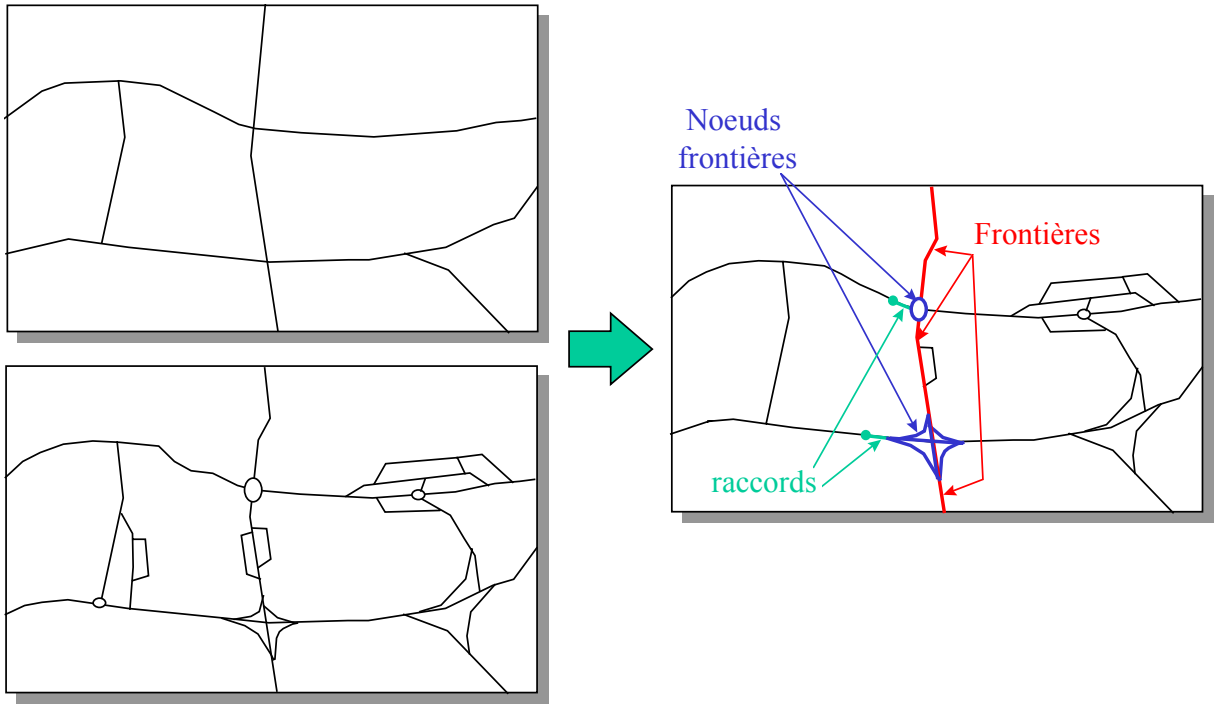



figure 86 : Exemple de dérivation de BDG hétérogène

Ces contraintes permettent de garder un graphe avec une représentation géométrique satisfaisante et de **dériver automatiquement une BDG hétérogène**. D'autres règles peuvent être définies, par exemple, sur les partitions de l'espace pour les objets surfaciques.

### *6.2.3 Perspectives pour des BDG fédérées*

Le processus d'intégration et d'appariement a été développé pour constituer des BDG intégrées sur un site central. Or, il n'est pas toujours possible de migrer les données des BDG sur un site unique. Plus précisément, s'il existe déjà des traitements importants sur des BD réparties, un système de **BD fédérées** (2.2.4.2) semble être la meilleure solution.

Le processus d'intégration et d'appariement doit donc être adapté, pour permettre la conception des BDG réparties. Dans cet objectif, les mécanismes d'appariement et l'emploi des règles de traduction doivent être modifiés.

#### 6.2.3.1 Modification du mécanisme d'appariement pour des BDG réparties

Le processus développé pour les BDG centralisées, apparie toutes les données de la base. Les données des BDG réparties évoluent indépendamment. Les appariements ne peuvent donc pas être calculés définitivement. L'appariement global doit être remplacé par un **appariement ciblé** exécuté lors de chaque requête.

Par exemple, pour une requête portant sur des tronçons $T_{1i}$ de la BD1 et $T_{2j}$ de la BD2, le processus d'appariement peut être le suivant :

- si les tronçons $T_{1i}$ et $T_{2j}$ appartiennent à des routes, ces routes sont appariées,
- un pré-appariement géographique des $T_{1i}$ et des $T_{2j}$ est réalisé,
- les noeuds extrémités des tronçons sont appariés,
- en fonction de l'appariement des routes, du pré-appariement des tronçons et de l'appariement des noeuds extrémités, un appariement des $T_{1i}$ et des $T_{2j}$ est réalisé à l'aide de filtres de plus court chemin entre les extrémités.

Pour les BDG réparties, cet appariement doit être entièrement fiable, rapide et automatique car une reprise manuelle n'est pas envisageable et le temps d'attente doit être court. De plus, les données réparties doivent être cohérentes. Ce n'est malheureusement pas le cas actuellement.

Deux **options** sont envisageables :

- les opérations sur les BDG réparties sont surtout des **consultations**. Les liens de correspondances peuvent alors être stockés localement ou sur un site distant. Lors d'une mise à jour, les liens sont soit recalculés, soit marqués comme étant obsolètes. Dans ce dernier cas, le lien est recalculé lors de la première requête l'utilisant.
- Cette solution n'est cependant pas satisfaisante du point de vue de la **cohérence** entre les BDG réparties et les liens de correspondances notamment pour les BDG ayant un nombre de mises à jour non négligeable. Pour conserver la cohérence, les BDG doivent utiliser des mécanismes de versionnement [Bauzer Medeiros et Jomier **93]** [Cellary et Jomier 90]. Ainsi, deux **versions** « stables » de la BDG pourront être appariées. Les liens de correspondance calculés et stockés entre les deux versions seront cohérents du fait du contrôle de cohérence inclus dans l'appariement. Pour mettre à jour ces versions, les utilisateurs dériveront de nouvelles versions, qui à leur tour, pourront être appariées.

#### 6.2.3.2 Règles de traduction automatiques

Pour une BD intégrée centralisée ou pour les BD réparties, des règles de traduction automatiques seront utilisées :

- pour migrer les données dans la BD centralisée,
- pour traduire les données dans le schéma de la BD répartie.

Par contre, pour les BD fédérées, les règles de traduction sont utilisées lors de chaque requête sur la base intégrée. En effet, si l'utilisateur pose sa requête sur le schéma fédéré, le processeur de requêtes réparties va scinder cette requête globale en un ensemble de sous-requêtes locales qui seront exécutées sur un site distant, puis l'ensemble des réponses sera récupéré pour constituer la réponse à la requête. Les règles de traduction sont donc utilisées lors de la définition des sous-requêtes et lors de la récupération des résultats.

Actuellement, pour migrer les données des règles de traduction manuelles ont été définies. il faut donc **définir des règles de traduction automatiques**. Cette tâche est complexe car les différences entre les schémas initiaux et le schéma intégré peuvent être nombreuses et compliquées. De plus, les systèmes de gestion des BD réparties peuvent être différents, ces systèmes hétérogènes nécessitent de définir des interfaces, pour autoriser la communication entre ces SGBD (interopérabilité entre SGBD). Les mécanismes de transfert définis lors de la pré-intégration devront de même être automatisés.

# 7. Annexes

## 7.1 Vocabulaire

Cette thèse emploie le vocabulaire issu de trois domaines :

- la modélisation conceptuelle orientée objet,
- la théorie des graphes
- les bases de données géographiques.

### *7.1.1 Vocabulaire de la modélisation conceptuelle orientée objet*

Les concepts orientés objet n'étant pas normalisés il faut donc les spécifier. Ces concepts sont fortement inspirés de la modélisation **OMT** [Rumbaugh et al. 91] et **UML** [Booch et al. 97]. Ces modélisations ont été préférées à ODMG [Atwood et al. 93] et EXPRESS [ISO 94] (STEP), qui se situent d'avantage à un niveau logique.

#### 7.1.1.1 Objets et classes

Un **objet** est une abstraction d'un phénomène du monde réel. Les objets sont les **instances** d'une classe.

Une **classe** rassemble un ensemble d'objets ayant les mêmes propriétés (attributs et méthodes)**.** Elle fournit une représentation commune décrite par des attributs et définie des comportements analogues représentés par des méthodes. Chaque objet connaît sa classe, et a un **identifiant** (OID), qui le distingue des autres objets et qui le référence.

Les **attributs** sont les variables des classes. Ils modélisent les caractéristiques des objets de la classe. Ces valeurs définissent l'état de l'objet. Chaque attribut a un **nom** unique pour sa classe et un **type** (type de valeur). Ce type peut être **atomique** (réel, entier, caractère, chaîne de caractères, booléen, énuméré[20]) ou **structuré** (n-uplet[21], ensemble[22] (set), liste[23] (list)). Un attribut doit être une pure valeur, il n'a pas d'identifiant. Un attribut peur être **optionnel**. Ainsi les caractéristiques inconnues, sans objet, ou n'existant pas pour certain des objets de la classe, sont possibles.

Les **méthodes** sont des fonctions qui peuvent s'appliquer à tous les objets de la classe. Les méthodes ont parmi leurs arguments, un argument implicite qui est l'objet cible de la classe appelant la méthode. Elles permettent de manipuler les objets. Une méthode est l'implémentation d'une **opération** pour une classe. La même opération peut être appliquée dans différentes classes. La méthode utilisée pour une opération dépend seulement de la classe de l'objet cible. Toutes les méthodes d'une opération ont la même signature (paramètres et résultat). Les attributs et les méthodes sont appelés les **propriétés** de la classe.

La relation de **sous-typage** entre les classes définit la substitution et la réutilisation autorisées entre les classes : si U est un sous-type de T alors les instances de U sont des instances de T (substitution) et de plus, U hérite de T, c'est-à-dire que les propriétés de T sont réutilisées pour

---

[20] Une **énumération** est un ensemble fini de libellés (chaîne de caractère le plus souvent), par exemple, couleur : enum :('bleu', 'vert', 'rouge').

[21] Un **n-uplet** est un ensemble finie d'attributs, par exemple, adresse : n-uplet (rue : string, numéro :integer, ville :string).

[22] Un **ensemble** est une collection de valeurs de même type, par exemple numéros_télephone : set (integer).

[23] Une **liste** est une collection ordonnée de valeurs du même type, par exemple, adresse_succesive : liste (adresse).

U (réutilisation). Ainsi défini le sous-typage inclus la notion d'héritage (réutilisation). Héritage et sous-typage seront donc utilisé indifféremment.

L'ensemble des objets d'une classe est appelé la **population** de la classe ou son **extension**. La population d'une classe peut être définie par son extension, en citant l'ensemble des instances ou par sont **intension**, en décrivant les instances de cette classe. Par exemple, la définition de la population de la classe zone boisée en intension est : espace peuplé d'arbres d'essence forestière (feuillus, conifères) dont la plus petite dimension est supérieure à 25 mètres. Cette définition en intension est appelée **spécification de contenu** pour les BDG.

#### 7.1.1.2 Liens et relations

Un **lien** est une connexion entre des objets. Par exemple, Jean Dupont *travaille pour* Industrie et Compagnie.

Une **relation** décrit un ensemble de liens ayant une structure commune et la même sémantique. Tous les liens d'une relation connectent des objets des mêmes classes. Un lien n'est pas une partie d'un objet, mais dépend d'au moins deux objets. Une relation peut avoir un nom (obligatoire en cas d'ambiguïté, deux relations reliant les mêmes types). Si une relation relie deux classes elle est dite **binaire**, si elle relie plus de deux classes, elle est dite **n-aire**. Pour chaque classe, les **cardinalités** d'une relation décrivent le nombre d'instances peuvent prendre part dans un lien de cette relation au minimum et au maximum. Une relation peut être décrite par des **attributs.** Leur définition est similaire à celle des attributs d'une classe. Les noms des **rôles** des classes dans la relation peuvent être ajoutés (obligatoire en cas d'ambiguïté, relation réflexive). Les objets d'un type participant à une relation peuvent être **ordonnés**. Les relations ne sont pas des classes, elles n'ont pas de méthodes, de relations de sous-typage et leurs instances n'ont pas d'identifiant.

### *7.1.2 Vocabulaire de la théorie des graphes*

Le modèle de graphe [Berge 83] est basé sur les concepts de **sommets** ou noeuds (vertex or node) V et **d'arêtes** ou arcs (edge). Un graphe est un couple (V,E) où V est l'ensemble des sommets et E un sous ensemble du produit cartésien $V \times V$. Un graphe est considéré **orienté** si l'ordre entre le sommet initial et le sommet final est important.

Les propriétés suivantes peuvent êtres définies sur les sommets :
- le **degré** d'un sommet est le nombre d'arêtes qui partent de ce sommet,
- le **degré** intérieur d'un sommet est le nombre d'arêtes qui entrent de ce sommet,
- le **degré** extérieur d'un sommet est le nombre d'arêtes qui sortent de ce sommet,
- les **voisins** d'un sommet sont tous les sommets reliés par des arêtes.

Pour un graphe les propriétés suivantes sont définies :
- un **cycle** d'un graphe est une suite de sommets partant d'un sommet et revenant sur ce sommet,
- un **chemin** est une suite d'arêtes successives qui ne forme pas un cycle,
- un graphe est **connexe** si chaque couple est lié par un chemin,
- un graphe est **biconvexe** si pour chaque paire de sommets i, j il y a un chemin de i à j et de j à i,

#### 7.1.2.1 Algorithme du plus court chemin

Nous allons décrire l'algorithme de plus court chemin de **Dijkstra** [Dijkstra 59] qui donne le plus court chemin d'un sommet V1 vers les autres sommets du graphe (V-). Il utilise un graphe (V,E) avec une fonction de coût toujours positive, par exemple une distance Dist(Vi,Vj).

**Initialisation**

$V- = V - V1$

$\forall Vi \in V-$

- si $V1 \times Vi \in E$ alors Dist_min (V1, Vi) = Dist (V1, Vi)
- sinon Dist_min (V1,Vi) = $\infty$

**Itération**

tant que $V- \neq \varnothing$

**1°) Sélection du sommet Vj non parcouru le plus proche de V1**

$Vj \in V- / \forall Vi \in V-$, Dist_min (V1, Vj) $\leq$ Dist_min (V1, Vi)

$V- = V- - \{Vj\}$

**2°) Pour les $Vi \in V-$ remplacer la distance minimum par la distance du chemin qui passe par Vj si il est plus court**

$\forall Vi \in V- / Vj \times Vi \in E$

si Dist_min (V1, Vj) + Dist (Vj, Vi) < Dist_min (V1, Vi)

alors Dist_min (V1, Vi) = Dist_min (V1, Vj) + Dist (Vj, Vi)

La théorie des graphes est utile pour modéliser un réseau. Un sommet est utilisé pour modéliser un noeud du réseau (par exemple un carrefour, une ville) et une arête est utilisée pour représenter un tronçon de ce réseau.

### *7.1.3 Vocabulaire de bases de données géographiques vectorielles*

Une **base de données géographique (BDG) vectorielle** décrit un ensemble de phénomènes du monde réel localisé sur la terre, organisée conformément au modèle conceptuel de données (MCD) géographiques afin de répondre à des besoins identifiés d'un ensemble d'utilisateurs. Cette base est stockée selon le format de données d'un SIG et est alors appelée **jeux de données**. Le passage peut nécessiter des enrichissements ou engendrer des pertes d'informations. Dans cette thèse, nous sommes restés à un niveau conceptuel, les problèmes liés au modèle physique ne seront que brièvement étudiés. Nous resterons donc au niveau des BDG.

Dans une BDG, l'information est décomposée en objets géographiques encore appelés entités géographiques (par exemple : des tronçons de route, des communes,…) et en liens entre ces entités (par exemple, une commune est responsable de l'entretien d'un tronçon de route). Ces entités et ces liens sont décrits suivant quatre niveaux :

- le **niveau sémantique** dans lequel les objets similaires sont regroupés en classes et les liens similaires en relations comme dans une BD classique,
- le **niveau topologique**, qui décrit les relations ou contraintes de proximité entre objets,

- le **niveau géométrique**, qui localise les objets par des coordonnées,
- le **niveau géodésique**, finalement, qui définit sans ambiguïté une position sur la Terre à partir des coordonnées.

#### 7.1.3.1 Le niveau géodésique

La localisation de l'information géographique se rapporte toujours à des coordonnées (couple ou triplet de valeurs numériques). Pour que ces coordonnées définissent sans ambiguïté une position sur la terre, il est nécessaire de leur associer une référence de coordonnées. Il existe trois méthodes principales pour associer une position sur la terre à des coordonnées [Rouet 91] :

- Les **coordonnées cartésiennes** (X,Y,Z) sont définies dans un référentiel géodésique qui est un repère affine défini par son origine O proche du centre de la terre, et trois axes orthogonaux (O, i), (O, j) et (O, k) avec (O, k) proche de l'axe des pôles, (O, i) proche du plan du méridien de Greenwich et (O, j) tel que (O, i, j, k) soit un repère direct.
- Les **coordonnées géographiques** (longitude, latitude et hauteur au-dessus de l'ellipsoïde) sont définies par un référentiel géodésique, un ellipsoïde géodésique et un méridien d'origine ainsi qu'une unité angulaire pour la longitude et la latitude et une unité linéaire pour la hauteur.
- Les **coordonnées cartographiques** (E, N, h) sont définies par des coordonnées géographiques et une projection cartographique qui est une transformation mathématique de la longitude et de la latitude définies sur un ellipsoïde vers les coordonnées cartésiens (E, N) définie dans un plan. h est la hauteur au-dessus de l'ellipsoïde.

#### 7.1.3.2 Le niveau géométrique

Le niveau géométrique regroupe les **primitives géométriques** qui sont construites à l'aide des coordonnées définie par le niveau géodésique. Ces primitives sont regroupées en **couches géométriques** auxquelles sont associées une référence de coordonnées et la résolution des coordonnées.

Les différentes primitives géométriques sont les suivantes :

- le **point**, décrit par ces coordonnées définissant ainsi la localisation,
- la **ligne**, décrite par une liste de points ainsi que d'éventuels paramètres d'interpolation entre deux points successifs. Par défaut, la localisation de la ligne est définie par des segments de droite joignant deux points successifs de la liste. Une ligne est dite **simple** si sa projection verticale ne s'intersecte pas avec elle même. Une ligne est dite fermée si les points extrêmes sont confondus,
- la **surface**, décrite par un contour extérieur et éventuellement des contours intérieurs appelés trous de la surface. Chaque contour est décrit par une ligne simple fermée.

#### 7.1.3.3 Le niveau topologique

Le niveau topologique est spécifié par la définition de couches topologiques qui sont, par définition, un ensemble de contraintes topologiques appliquées à un ensemble d'objets.

Les **contraintes topologiques** les plus fréquentes sont :

- la contrainte **d'identité** de la géométrie sous certaine condition (si un tronçon longe un bois alors la localisation du tronçon et d'une partie du contour du bois doit être identique),
- la contrainte de **non intersection** entre géométries des objets d'une même classe (les tronçons de route ne doivent pas s'intersecter sauf au carrefour),

- la contrainte de **partition** formée par un ensemble de surfaces (pour la France, les surfaces des communes forme une partition).

Dans ce niveau des **relations topologiques** peuvent aussi apparaître.

Ces relations topologiques sont des relations pré-définies obligatoires formant un modèle topologique. Plusieurs modèles ont été définis. Ils peuvent être groupés en deux familles :

- les modèles topologiques de **graphe** ou de **réseau** : Ces modèles obligent les arcs à avoir obligatoirement un, et un seul noeud initial et un et un seul noeud final. Inversement, un noeud peut être relié à aucun, un ou plusieurs arcs en tant que noeud initial ou noeud final. Deux sous-modèles existent le modèle **non planaire** et le modèle topologique de graphe **planaire,** qui de plus oblige la présence d'un noeud lors du croisement de deux arcs.
- les modèles topologiques de **carte** ou de **surface** (figure 87) sont plus complets. Ils ajoutent aux modèles topologiques de réseau planaire, la prise en compte obligatoire, des surfaces délimitées à gauche et à droite de chaque arc. Un arc a donc une et une seule surface à droite et une et une seule surface à gauche. Inversement une surface est située soit à gauche, soit à droite d'un ou plusieurs arcs.

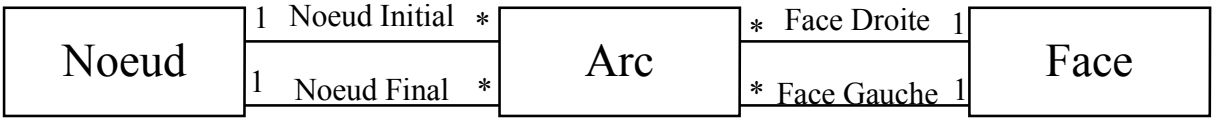



figure 87 : Modèle topologique de carte ou de surface

Pour une BDG, un modèle topologique n'est pas obligatoire, une BDG sans topologie est dite **spaghetti**. La topologie est un concept fondamental pour les SIG. En effet, elle fait partie de la gestion de la cohérence pour les BDG et facilite les requêtes spatiales. Cependant, c'est un mécanisme de gestion coûteux et elle augmente le nombre de géométrie.

#### 7.1.3.4 Le niveau sémantique

Le niveau sémantique décrit le découpage de l'information en objets géographiques ainsi que les caractéristiques associées à chacun des objets. Ce découpage est régi par des contraintes topologiques définies au niveau topologiques et par des principes d'homogénéité des objets géographiques. A un objet géographique est associé un ensemble de valeurs décrivant ses caractéristiques (comme le nom d'une commune ou la largeur d'un tronçon de route) et chaque objet géographique doit être « **homogène** » pour chacune de ses caractéristiques. Cette homogénéité est modulée par une **granularité** qui limite les parties hétérogènes dans l'espace (par exemple, une partie plus étroite du tronçon courte ne provoquera pas de découpage, si sa longueur est inférieure à la granularité définie pour la largeur).

Les objets ayant des caractéristiques semblables, sont regroupés en **classes** et chaque caractéristique est modélisée par un **attribut** associé à la classe. Des **relations** entre classes décrivent les liens possibles entre objets géographiques.
L'information géographique peut souvent être décomposée hiérarchiquement en objets qui sont eux-mêmes des ensembles d'objets. L'exemple le plus typique est la décomposition hiérarchique en unités administratives, une région est composée d'un ensemble de départements, chacun composé d'un ensemble de cantons. Pour simplifier la modélisation, on distingue des autres relations, la relation de **composition (**relation d'agrégation) qui sont des associations qui décrivent un objet d'une classe par un ensemble d'objets d'autres classes.

Un objet géographique est localisé à l'aide des primitives géométriques.

## Notions d'Unified Modeling Language (UML)

UML (Unified Modeling Language) [Booch et al. 97] est une méthode pour spécifier, visualiser et documenter l'artefact d'un système orienté objet en phase de développement. UML est issue de l'unification des méthodes de **Booch** [Booch 91] et de **OMT** [Rumbaugh et al. 91] et doit donc s'imposer comme un standard.

Nous allons juste décrire les symboles graphiques nécessaires à la compréhension des schémas conceptuels (diagrammes de classes).

Une **classe** est représenté par un rectangle avec trois composants, le non de la classe en haut, une liste d'attributs avec le type de l'attribut (optionnel) au milieu et la liste des opérations en bas (nous ne traiterons pas des opérations).

Une **relation binaire simple** est représentée par une ligne entre les deux classes. Le nom de la relation est écrit à coté de la ligne. Chaque extrémité de la relation a un rôle qui peut avoir un nom. Chaque rôle indique la cardinalité de sa classe (c'est-à-dire combien d'instances de cette classe peuvent être associées avec une instance de la classe en relation). La cardinalité est exprimé par un intervalle (valeur minimale, deux points, valeur maximale), * indique que le nombre d'objet est illimité. Si la cardinalité est 0..* la cardinalité est exprimée par *. Le mot clé {ordered} peut être placé dans le rôle, il indique que les éléments ont un ordre explicite.

Les **relations réflexives ou n-aire** sont aussi possibles, le nom des rôles sont alors obligatoires. Les relations n-aires sont représentées par des lignes reliant chaque rôle à un losange.

Les relations peuvent avoir des attributs et des opérations, dans ce cas une **classe relation** est dessinée. Elle est représentée par un rectangle avec les trois mêmes composants que la classe, et est reliée à la ligne de la relation (ou le losange pour les relations n-aire) par une ligne en pointillé.

Une **relation de composition** est une relation spécifique avec une connotation « composé - composant » (« whole - part »). Elle est représenté par un losange du coté du composé.

L'**héritage** entre une super classe et une sous classe est représenté par une flèche orientée vers la super classe.

## 7.2 Autres applications multi-représentations

### *7.2.1 Autres exemples de cartes électroniques multi-représentations*

Le projet Américain « **Alexandria Digital Library** » [Frew et al. 95] qui consiste à créer une bibliothèque électronique sur les informations géographiquement référencées réparties. Cette bibliothèque dispose de différentes cartes électroniques à différentes échelles, et d'une interface qui permet de changer de cartes.

Des Atlas sur CD ROM tel **Encarta world Atlas** de Microsoft® qui permet de disposer de plusieurs représentations à des échelles différentes sur une même zone.

Le site internet **Interactive Atlas Map Quest** [GeoSystems Global Corp 96] permet d'afficher des cartes du monde entier à différentes échelles allant environ du 1 :1 000 000 000 pour visualiser le monde au 1 :10 000 pour visualiser les rues d'une ville.

Le site web **Multi-Scale Maps** de l'université de Californie [Bradley 96] permet de visualiser des cartes rasters multi-échelles.

**GeoKiosk** [ESRI France 97] est un nouvel outil d'observation de données géographiques. L'utilisateur survole librement son espace géographique à n'importe quelle altitude et examine toujours l'information la plus adaptée à son regard. La première application réalisée avec le produit GeoKiosk, est une borne de consultation grand public du canton de Genève à partir de cartes scannées, d'un plan de ville, et d'une base de données vecteurs contenant environ 70 000 parcelles.

### *7.2.2 Autres exemples d'analyse multi-représentation*

Analyse multi-échelle des structures de l'occupation du sol au Canada [Fraser 95],

Analyse multi-échelle de la répartition des nuisances à Houston [Sui et Giardino 95],

Simulation multi-échelle des flux de carbone dans les zones humides boisé de l'Ontario [Band. 94]

Analyse multi-échelle de la gestion de l'écosystème [Bennett et Sharpe 95]

Analyse multi-échelle des ratios de vente de propriétés [Noonan, et Cowen 95]

Evaluation multi-échelle du réseau fluviale de la région de forêt des grands lacs [Perera et al. 94].

Analyse statistique multi-échelle des l'îlots, le logement, le ménage et la personne [Piron 92] [Piron 93]

Analyse multi-échelle statistique de l'impact du remplacement de la Poll Tax par la Council Tax à Cardiff, au niveau du quartier, de la rue, de la propriété et selon le type d'habitation [Higgs et al.94].

Le serveur régional Midi-Pyrénées, d'information géographiques qui permet des analyse multi-échelle [Crépeau et Huet 93].

### *7.2.3 Autres exemples de contrôle de cohérence*

Contrôle de cohérence des données géographiques sur l'environnement en Ecosse [Brooker 95].

Comparaison des frontières des bassins hydrographiques à différentes échelles aux Etats-Unis [Peltz et Liebermann 95]

Comparaison de la BD TOPO [IGN 96 b] (4.1.1.1) et du Cadastre en France [Lemarié 96].

### *7.2.4 Exemples de plate-formes d'intégration pour le travail coopératif*

plate-forme de programmes scientifiques multi-disciplinaires [Alm et al. 94]

plate-forme de gestion de l'évolution des zones urbaines [Lahti 97]

plate-forme d'analyse écologique [Karra et al. 94]).

## 7.3 Opérations d'intégration

Les opérations d'intégration, doivent répondre à deux problèmes, la **mise en conformité** des classes et de leur propriété, et **l'intégration de ces classes**. Motro [Motro 87] a défini 10 opérations d'intégration (figure 21 page 3). Nous allons présenter les 15 opérations d'intégration proposées par Dupont [Dupont 95 b].

### *7.3.1 Les opérations d'intégration de classe*

[Dupont 95 b] a déterminé 15 opérations d'intégrations qui sont utilisées pour relier les classes à intégrer. Le résultat peut être un ensemble de classes simples ou des classes plus complexes s'appuyant sur les concepts de généralisation - spécialisation, ou de multi-instanciation. Ces 15 opérations vont être présentées maintenant, en commençant par les opérations les plus simples. Pour les illustrer, l'intégration de l'exemple suivant sera réalisée avec chacune des opérations.

| Classe **Port de Plaisance** | Classe **Port de Pêche** |
|---|---|
| toponyme : chaîne de caractères | toponyme : chaîne de caractères |
| nb d'emplacements : entier | tonnage : entier |

#### 7.3.1.1 La préservation

La préservation n'est pas à proprement parler une opération d'intégration, mais plutôt une technique possible. Elle consiste à ne rien changer. Pour l'exemple, le résultat de cette opération est donc l'exemple.

| **Port de Plaisance** |
|---|
| toponyme |
| nb emplacements |

| **Port de Pêche** |
|---|
| toponyme |
| tonnage |

#### 7.3.1.2 La fusion

La fusion consiste à créer dans le schéma intégré, une classe ayant pour attributs l'ensemble des attributs des classes à intégrer et pour instances l'union des instances. L'ensemble des phénomènes du monde réel représenté dans l'une des classes à intégrer est donc représenté dans la classe résultante. Les propriétés spécifiques aux classes à intégrer, deviennent facultatives dans la classe intégrée. Cette opération est équivalente à l'opération Combine de Motro.

Pour l'exemple, la fusion crée une classe **Port** ayant pour instances les ports de plaisance et les ports de pêche, pour attribut obligatoire le toponyme et pour attributs spécifiques nb emplacements et tonnage.

| **Port** |
|---|
| toponyme |
| nb emplacements |
| tonnage |

Cette opération est utile si les classes sont très proches, elle permet d'obtenir un schéma intégré simple conservant l'ensemble des informations.

#### 7.3.1.3 L'union

L'union consiste à créer dans le schéma intégré, une classe unique ayant pour attributs les attributs communs et pour instances l'union des instances.

Pour l'exemple, l'union crée une classe **Port** ayant pour instances les ports de plaisance et les ports de pêche et pour attribut obligatoire l'attribut commun : toponyme

| **Port** |
|---|
| toponyme |

Cette opération est proche de la fusion, hormis les attributs spécifiques qui ne sont pas conservés. Elle est utile si la base intégrée doit gérer uniquement les descriptions communes.

#### 7.3.1.4 L'intersection

L'intersection consiste à créer dans le schéma intégré, une classe unique ayant pour attributs l'ensemble des attributs des classes à intégrer et pour instances les intersections des instances.

Pour l'exemple, l'intersection crée une classe **Port Mixte** ayant pour instances les ports qui font à la fois partie de la classe port de plaisance et de la classe port de pêche. Cette classe a pour attribut obligatoire les trois attributs initiaux.

| **Port Mixte** |
|---|
| toponyme<br>nb emplacements<br>tonnage |

Cette opération est utile, si la base intégrée ne doit concerner que l'intersection des instances.

#### 7.3.1.5 La partition

La partition consiste à créer une classe pour chaque intersection et pour chaque différence. Les attributs des intersections sont l'union des attributs. Les attributs des différences sont les attributs de la classe d'origine.

Pour l'exemple, 3 classes sont crées dans le schéma intégré :

- la classe des **Ports exclusivement de Plaisance** avec pour attributs les attributs de la classe **Port de Plaisance**,
- la classe des **Ports exclusivement de Pêche**,
- la classe des **Ports Mixtes** qui regroupe les instances de l'intersection.

| **Port exclusivement de Plaisance** |
|---|
| toponyme<br>nb emplacements |

| **Port Mixte** |
|---|
| toponyme<br>nb emplacements<br>tonnage |

| **Port exclusivement de Pêche** |
|---|
| toponyme<br>tonnage |

Cette opération est intéressante, car elle permet de résoudre les problèmes d'intersection entre les classes à intégrer tout en conservant l'ensemble de l'information. Par contre, elle est difficilement utilisable pour plus de deux classes à intégrer, car elle produit un grand nombre de classes.

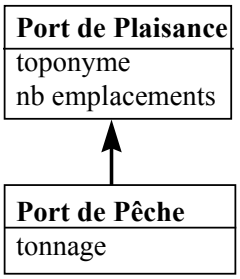
Les classes intégrées peuvent aussi utiliser les relations de **généralisation spécialisation**.

#### 7.3.1.6 La sous-classe

L'opération sous-classe consiste à définir une relation d'héritage entre les deux classes. Cette opération peut être uniquement utilisé si l'extension de la classe fille est inclus dans celle de la mère. Cette opération est équivalente à l'opération Connect de Motro.

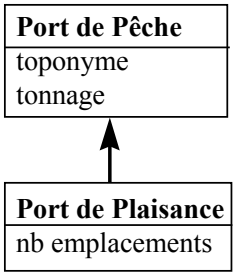
Pour l'exemple, si tous les ports de pêche sont des ports de plaisance, alors l'opération sous-classe peut être utilisée, le schéma intégré est alors composé d'une classe générique **Port de Plaisance** et d'une classe spécifique **Port de Pêche**. **Port de Plaisance**

a pour attributs ses attributs propres et pour instances les ports de plaisance qui ne sont pas des ports de pêche. **Port de Pêche** a pour attributs ses attributs spécifiques et pour instances les ports de pêche qui sont aussi des ports de plaisance.




#### 7.3.1.7 La sur-classe

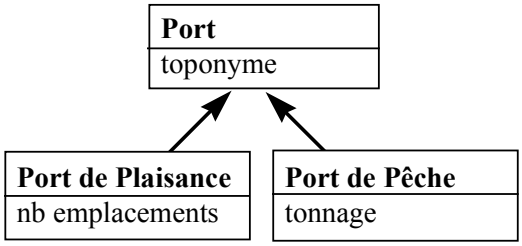
Cette opération est l'opération inverse de sous-classe. Cette opération est équivalente à l'opération Connect de Motro mais avec les paramètres inversés.




#### 7.3.1.8 La généralisation

L'opération de généralisation consiste à définir une classe générique, et des liens d'héritages entre les classes à intégrer et cette classe générique. Cette opération permet d'unifier les descriptions mais ne permet pas de gérer les intersections. Elle équivaut à l'opération Meet de Motro.

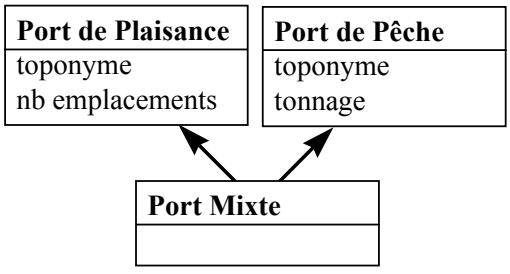
Pour l'exemple, une classe générique Port est crée et les deux classes à intégrer héritent de cette classe. La classe Port a pour attribut, toponyme qui est commun aux classes initiales.




#### 7.3.1.9 La spécialisation

L'opération de spécialisation consiste à définir une classe spécifique aux classes à intégrer. Cette technique permet de gérer proprement l'intersection. Cette opération est équivalente à l'opération Join de Motro.

Pour l'exemple, une classe spécifique **Port Mixte** est créée, elle hérite des classes **Port de plaisance** et **Port de pêche**, elle n'a pas d'attribut. Les instances de **Port Mixte** sont les instances de l'intersection **de Port de Plaisance** et **Port de Pêche**. Les classes mères ont donc pour instance la différence.



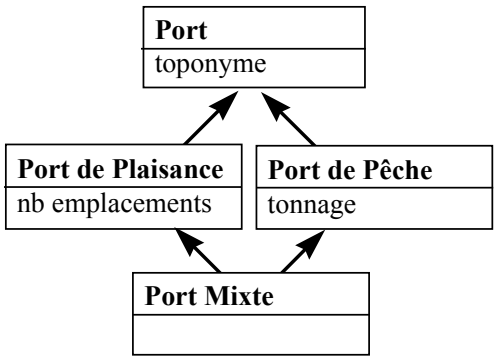

Les opérations suivantes sont des combinaisons de deux opérations.

#### 7.3.1.10 La généralisation - spécialisation

L'opération de généralisation - spécialisation est une combinaison des opérations de généralisation et de spécialisation. Elle permet de gérer proprement l'intersection et l'union des extensions, et de réduire les redondances au niveau des structures. Par contre, elle génère un grand nombre de classes et une hiérarchie multiple.

Pour l'exemple, 4 classes sont créées :

- la classe générique **Port** qui a pour attribut propre toponyme et pas d'instances propre,
- la classe **Port de Plaisance** qui hérite de **Port**, a pour attribut propre nb emplacement, et pour instances propre, les instances qui sont uniquement des ports de plaisance,
- la classe **Port de Pêche** qui hérite de **Port**, a pour instances propres les ports qui sont uniquement des ports de pêche, pour attribut propre tonnage,
- la classe **Port Mixte** qui hérite de **Port de Plaisance** et **Port de Pêche**, qui n'a pas d'attribut propre et pour instance les ports qui sont à la fois des ports de pêche et des ports de plaisance.

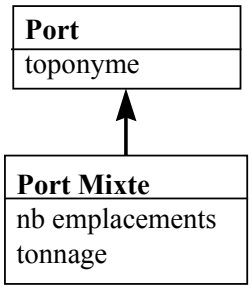



#### 7.3.1.11 Union - Intersection

L'opération d'union - intersection est une combinaison des opérations d'union et d'intersection. Elle permet de gérer à la fois l'union des instances et leurs intersections. Par contre, pour les instances sans correspondance, elle ne conserve pas l'information détenue par les attributs spécifiques.

Pour l'exemple, cette opération définie deux classes :

- La classe **Port** qui a pour attribut toponyme et pour instances les ports étant exclusivement de plaisance ou exclusivement de pêche.
- La classe **Port Mixte** qui à pour instances les ports mixtes et pour attribut nb emplacement et tonnage



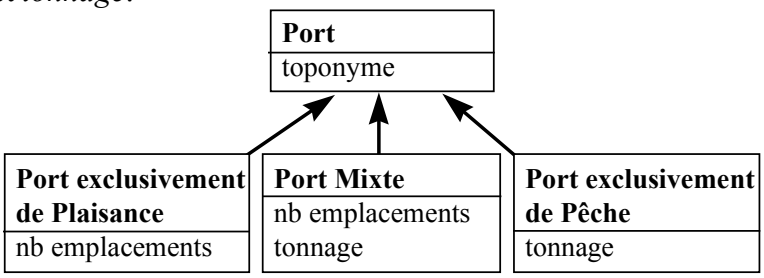

#### 7.3.1.12 L'union - partition

L'opération d'union - partition est une combinaison des opérations d'union et partition. Elle permet de disposer d'une classe générique (l'union) et de classe spécifique issue de la partition. Elle conserve l'ensemble des instances sans perte d'information.

Pour l'exemple, cette opération définie quatre classes ;

- la classe **Port** qui a pour attribut toponyme et qui n'a pas d'instance propre.

- la classe des ports exclusivement de plaisance avec pour attributs, l'attribut spécifique de la classe port de plaisance : *nb emplacements* et pour instances les ports qui sont uniquement de plaisance.
- la classe des port exclusivement de pêche, avec pour attributs, l'attribut spécifique de la classe port de pêche : *tonnage*, et pour instances les ports qui sont uniquement de pêche.
- la classe des ports mixtes qui regroupe les instances de l'intersection des deux classes à intégrer et qui a pour attributs l'ensemble des attributs spécifiques à ces deux classes : *nb emplacement* et *tonnage*.

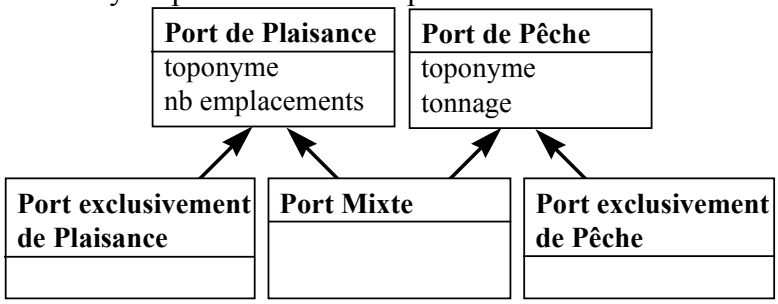



#### 7.3.1.13 la spécialisation - partition

L'opération de spécialisation - partition est une combinaison des opérations de spécialisation et de partition. Elle permet de gérer les différences et l'intersection tout en conservant les classes d'origine.

Pour l'exemple, cinq classes sont engendrées par cette opération :

- la classe **Port de Plaisance** ayant pour attributs toponyme et nb emplacement,
- la classe **Port de Pêche** ayant pour attributs toponyme et tonnage,
- la classe **Port exclusivement de Plaisance** ayant pour instances les ports exclusivement de plaisance,
- la classe **Port exclusivement de Pêche** ayant pour instances les ports exclusivement de pêche,
- la classe **Port Mixte** ayant pour instances les ports mixtes.



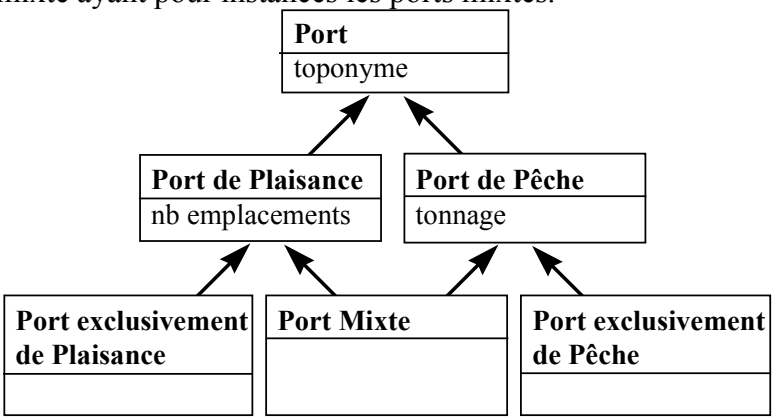

#### 7.3.1.14 La généralisation - partition

L'opération de généralisation - partition est une combinaison des opérations de généralisation et de partition. Elle permet de gérer l'union, l'intersection, les différences et les classes d'origine au prix d'une hiérarchie complexe.

Pour l'exemple, six classes sont engendrées par cette opération :

- la classe port ayant pour attribut toponyme,
- la classe port de plaisance ayant pour attribut nb emplacement,
- la classe port de pêche ayant pour attribut tonnage,
- la classe port exclusivement de plaisance ayant pour instances les ports exclusivement de plaisance,

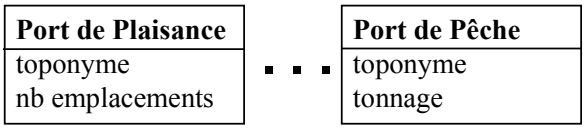
- la classe port de pêche ayant pour instances les ports exclusivement de pêche,
- la classe port mixte ayant pour instances les ports mixtes.




#### 7.3.1.15 La muti-instanciation

L'opération de muti-instanciation s'appuie sur le concept d'instanciation multiple. Dans les BD classiques un objet est l'instance d'une seule classe. Avec ce nouveau concept, un objet peut être l'instance de plusieurs classes non reliées par des liens d'héritage.

Pour l'exemple, la technique de muti-instanciation va permettre pour les ports mixtes, d'être des instances des classes port de plaisance et port de pêche.




### *7.3.2 Intégration des relations*

En ce qui concerne les relations, seules les cinq techniques d'intégration des classes n'utilisant pas le concept d'héritage et d'instanciation multiple, peuvent être utilisées :

- la **préservation** consiste simplement à placer les relations dans le schéma intégré sans aucune modification,
- la **fusion** crée une relation unique dans le schéma intégré. Elle regroupe les liens des deux relations et les attributs des deux relations. Les attributs propres deviennent optionnels,
- **l'union**. crée une relation unique par regroupement des instances des deux relations qui a pour attribut uniquement les attributs communs. Cette technique ne sera pas utilisée, car elle ne conserve pas les attributs propres.
- **l'intersection** crée une relation unique. Elle gère uniquement les liens ayant des correspondants et l'ensemble des attributs. Cette technique ne sera pas utilisée, car les liens propres ne sont pas conservés.
- la **partition** en relation complémentaire. Trois relations sont créées ; une gérant les liens communs, et deux gérant les différences. La première porte l'ensemble des attributs, les deux suivantes les attributs de leur relation d'origine.

## 7.4 Signature des outils d'appariement de la boîte à outils

### *7.4.1 L'outil sémantique*

| |
|---|
| **égalite_semantique** ( objets1 : set(objet_bd1), propriétés1 : list(string),<br>objets2 : set (objet_bd2), propriétés2 : list(string)) :<br>set (n-uplet ( appariés1 : set(objet_bd1),<br>appariés2 : set(objet_bd2))) |

### *7.4.2 Les outils géométriques de distance*

| |
|---|
| **objets ponctuels** |
| **Distance_euclidienne** (objet_ponctuel1 : objet_bd1, objet_ponctuel2 : objet_bd2) : réel |
| **objets linéaires** |
| **Distance_moyenne** (objets_linéaires1 : set(objet_bd1), objets_linéaires2 : set(objet_bd2)) : réel [McMaster 86] |
| **Distance_Fréchet_d** (objets_linéaires1 : set(objet_bd1), objets_linéaires2 : set(objet_bd2), précision : réel) : réel [Eiter et Mannila 94] |
| **Distance_Hausdorff_d** (objets_linéaires1 : set(objet_bd1), objets_linéaires2 : set(objet_bd2), précision : réel) : réel [Branly 97]. |
| **Appariement_composante_Hausdorff** (objets_à_comparer : set(objet_bd1), objets_fixés : set(objet_bd2), seuils : list(réel), pas : réel) : set (n-uplet (objet_comparé : objet_bd1, objets_fixés : set(objet_bd2)) |
| **objets surfacique** |
| **Distance_surfacique** (objet_surfacique1 : objet_bd1, objet_surfacique2 : objet_bd2) : réel [Vauglin 97] |

### *7.4.3 Les outils géométriques de forme*

| |
|---|
| **objets linéaires** |
| **rapport_longueurs_arcs** (objet_linéaire1 : objet_bd1, objet_linéaire2 : objet_bd2) : réel |
| **rapport_longueurs_moyennes_virages** (objet_linéaire1 : objet_bd1, objet_linéaire2 : objet_bd2) : réel |
| **différence_directions** (objet_linéaire1 : objet_bd1, objet_linéaire2 : objet_bd2) : réel |
| **rapport_nombre_ points_intermédiaires** (objet_linéaire1 : objet_bd1, objet_linéaire2 : objet_bd2) : réel |
| **rapport_sommes_angles** (objet_linéaire1 : objet_bd1, objet_linéaire2 : objet_bd2) : réel |
| **rapport_nombre_virages** (objet_linéaire1 : objet_bd1, objet_linéaire2 : objet_bd2) : réel |
| **objets surfaciques** |
| **rapport_aires** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel |
| **rapport_périmètres**(objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel |

**rapport _ompacités** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_élongations** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_excentricités** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_allongements** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_rectangularités** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_distance_maximum_centroïde** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_distances_radiales_centroïde** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : set (réel)

**rapport_nombr_trous** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_nombre_de_composants_connexes** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

**rapport_nombre_Euler** (objet_surfacique : objet_bd1, objet_surfacique : objet_bd2) : réel

### *7.4.4 Les outils topologiques*

**Plus_court_chemin** (sommets_départ : objet_bd2, sommets_arrivé : objet_bd2, graphe : set (objet_bd2)) : set (objet_bd2),

**Regroupement_connexe** (tronçons : set (objet_bd), noeuds : set (objet_bd)) : set(tronçons_du groupe : set (objet_bd), noeuds_du_groupe : set (objet_bd))

**Impasse** (arête : objet_bd1) : booléen

**nb_arêtes** (sommet : objet_bd1, propriétés : list(string), types : list(string), valeurs : list(string)) : entier

### *7.4.5 Les outils mixtes*

**Plus_proche_chemin** (sommet_depart : objet_bd2, sommet_arrivé : objet_bd2, graphe : set (objet_bd2), chemin : set(objet_bd1)) : set (objet_bd2),

## 7.5 Copies d'écran des résultats de l'appariement sur la zone de Marne-la-Vallée Lagny

### *7.5.1 Appariement géométrique*

2 copies d'écran illustrant les résultats de l'appariement géométrique à l'aide de la composante de la distance de Hausdorff par un seuillage successifs (sous section 5.1.1.2) des données routières de la BD CARTO (en gris) et des données de GEOROUTE (en couleur).

**légende**

**gris** : tronçon BD CARTO

**vert** : tronçon GEOROUTE apparié géométriquement
**jaune** : tronçon GEOROUTE non apparié géométriquement
**rouge** : tronçon GEOROUTE litigieux

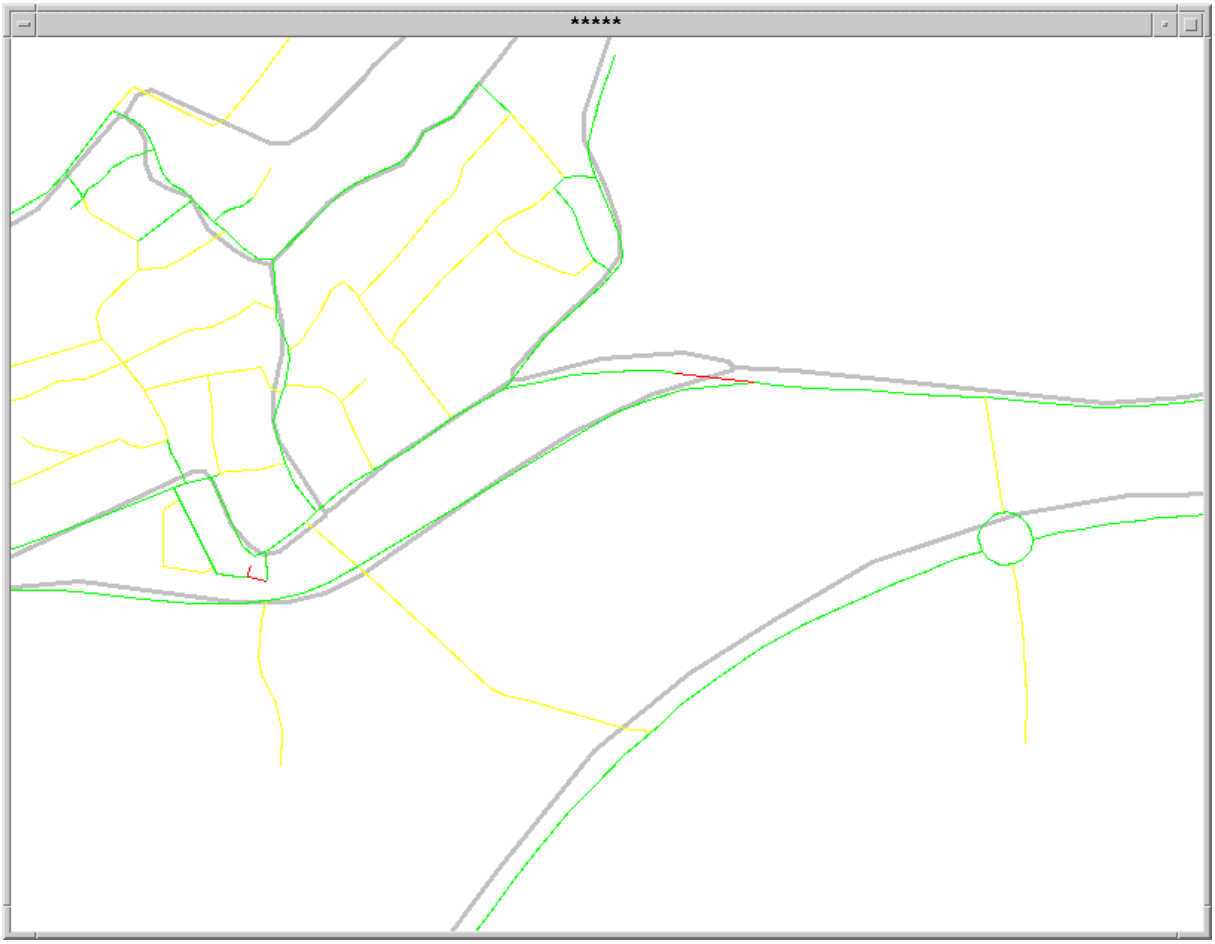



figure 88 : Exemple 1 d'appariement géométrique des tronçons


figure 89 : Exemple 2 d'appariement géométrique des tronçons dans une zone où les représentations sont relativement incohérentes

## *7.5.2 Appariement après filtrage*

copies d'écran des résultats de l'appariement après filtrage

**légende**

**gris** : tronçon BD CARTO apparié
**noir** : tronçon BD CARTO sans correspondant

**vert** : tronçon GEOROUTE apparié
**magenta** : tronçon GEOROUTE apparié géométriquement mais non topologiquement
**jaune** : tronçon GEOROUTE non apparié géométriquement
**bleu** : tronçon GEOROUTE apparié avec un noeud
**bleu ciel** : tronçon GEOROUTE apparié géométriquement avec un noeud mais non topologiquement
**rouge** : tronçon GEOROUTE litigieux

### 7.5.2.1 résultats sur la zone de Montévrain

Les données de la BD CARTO et de GEOROUTE de cette zone sont présentées dans la figure 61et la figure 62.

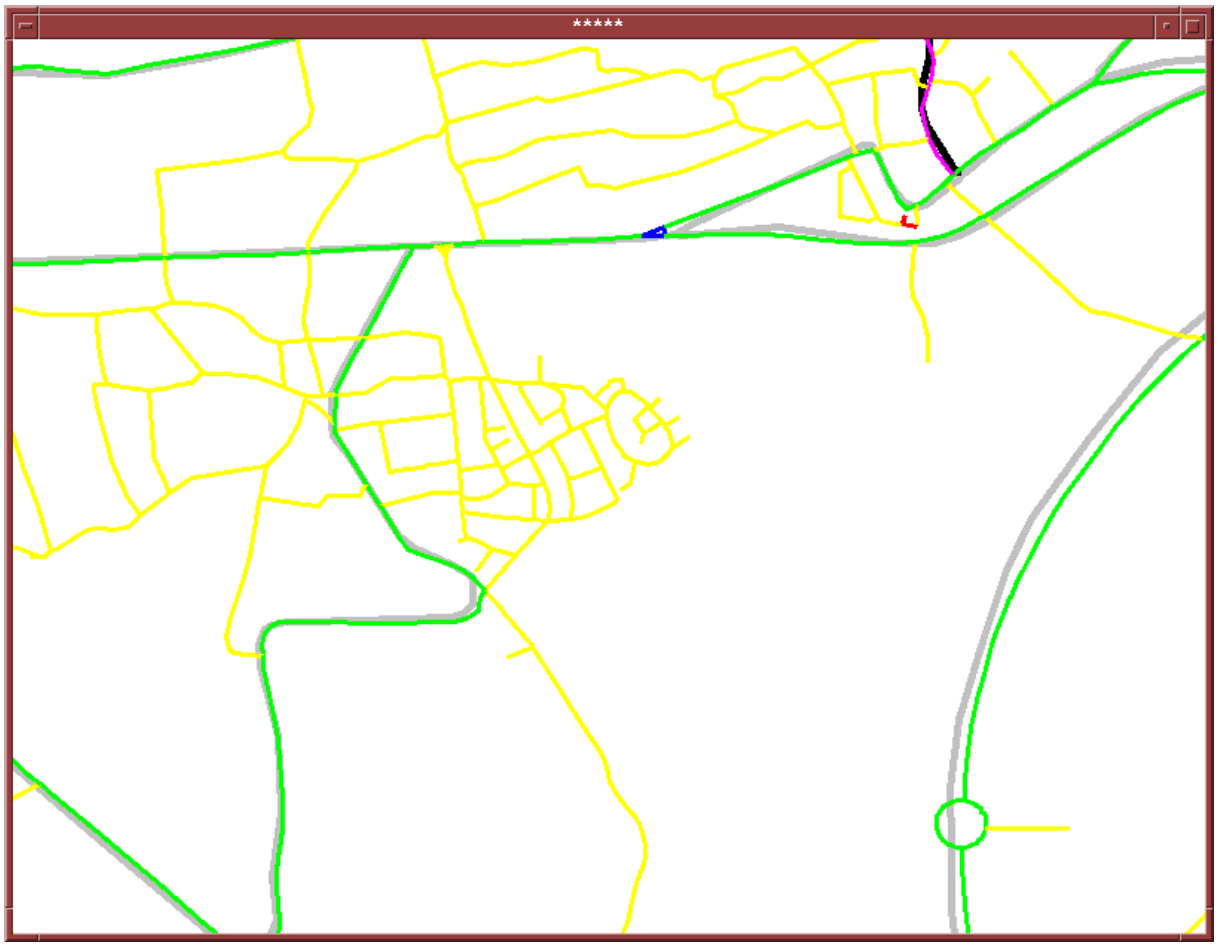



figure 90 : Résultat de l'appariement sur la zone de Montévrain

#### 7.5.2.2 Exemples d'appariements corrects

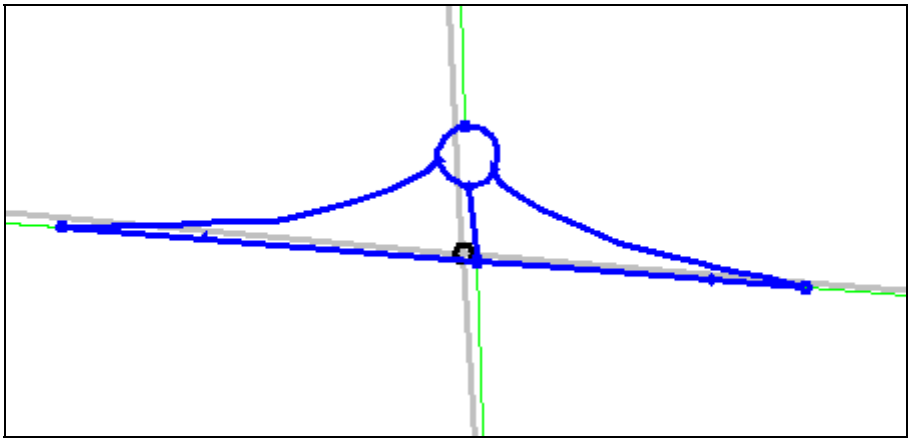


figure 91 : Exemple 1 d'appariement 1-n de noeuds routiers

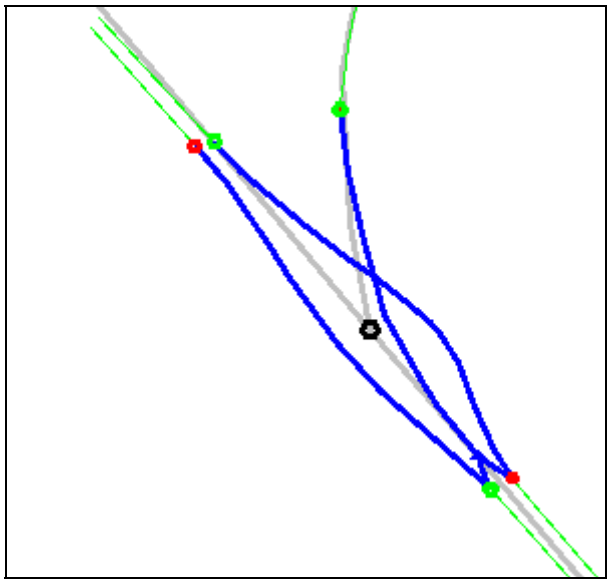


figure 92 : Exemple 2 d'appariement 1-n de noeuds routiers

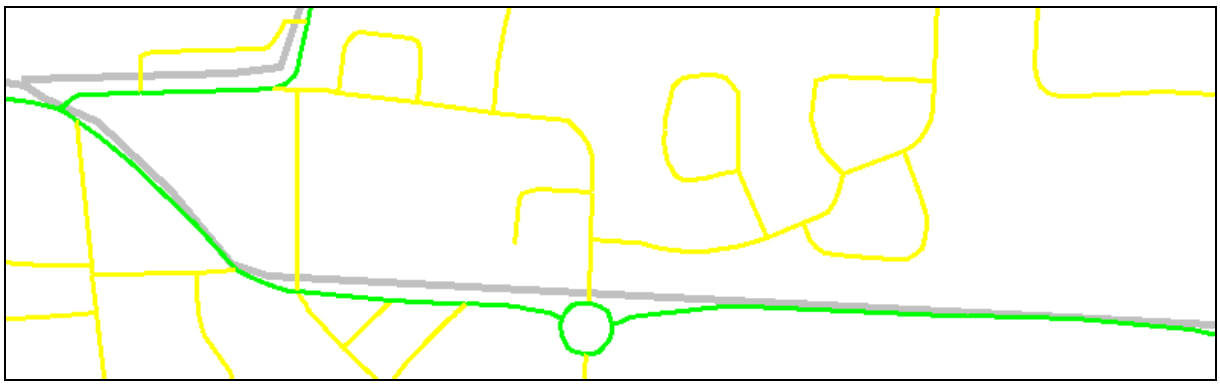


figure 93 : Exemple d'appariement 1-n de tronçons routiers

### 7.5.2.3 Exemples d'appariements impossibles où incorrects

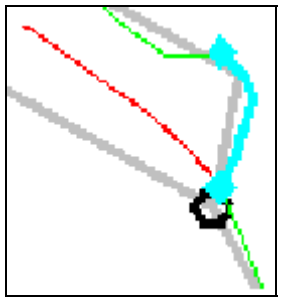


figure 94 : Exemple d'appariement géométrique d'un des tronçons sortant du groupe impossible (tronçon GEOROUTE entre 2 tronçons BD CARTO parallèle et proche)

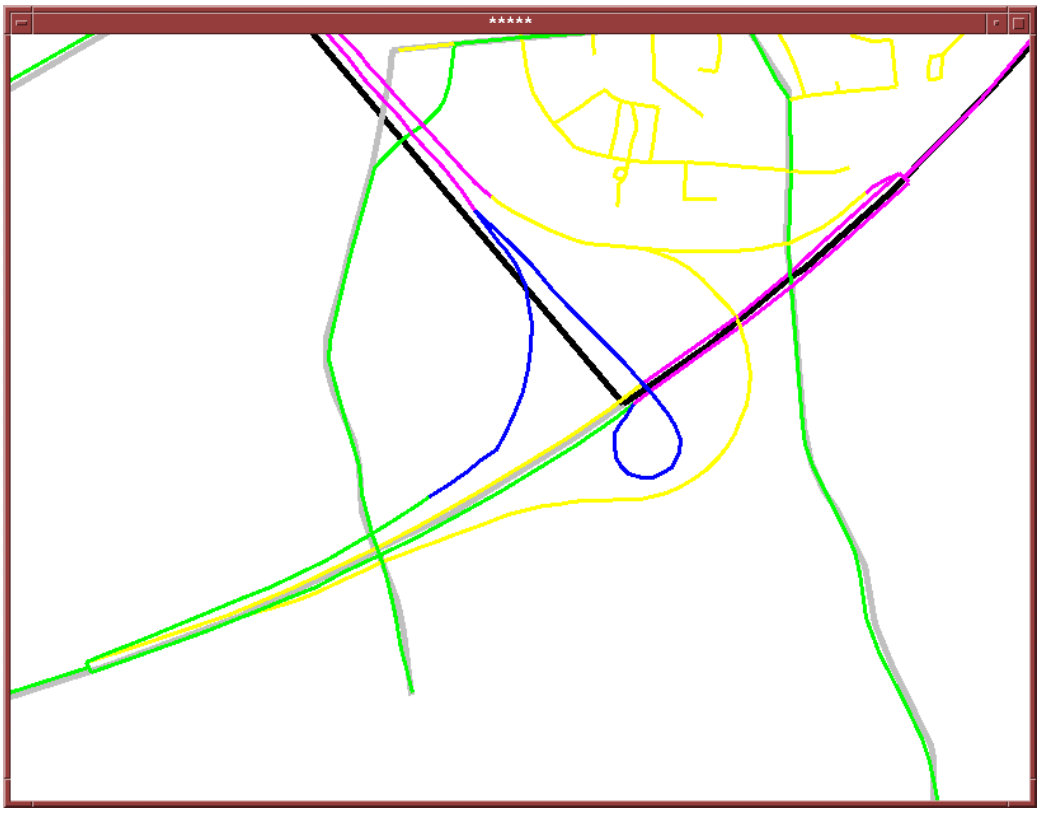



figure 95 : Exemple d'appariement impossible entre un noeud BD CARTO avec un échangeur non inclus dans la zone de recherche (∅ de l'échangeur 2,2 km)

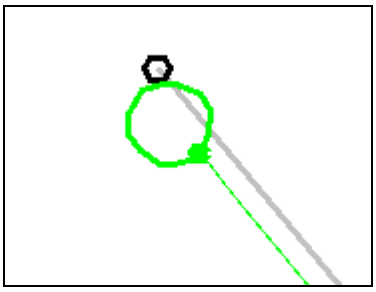


figure 96 : Exemple d'appariement 1-n des noeuds non détecté
(rond-point cul-de-sac)

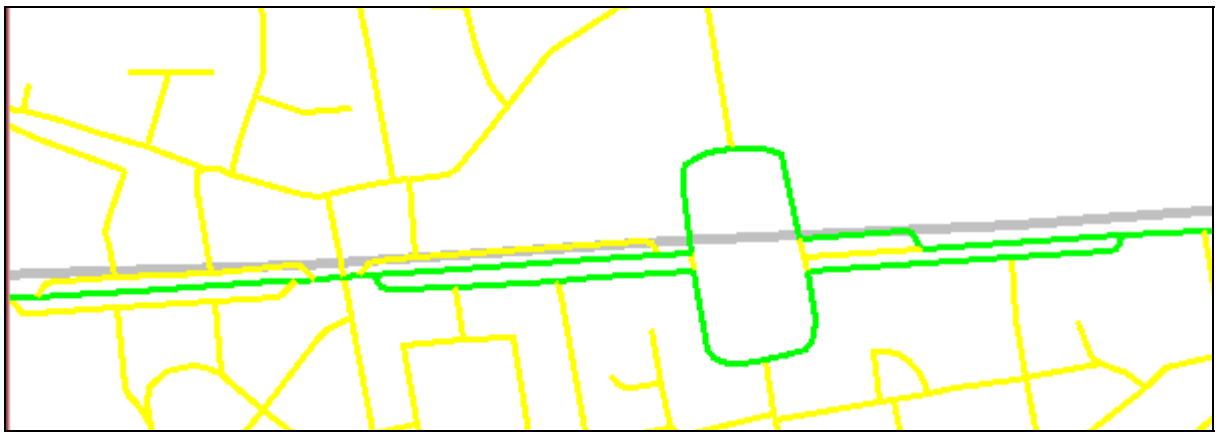


figure 97 : Exemple de mauvais choix des tronçons lors du filtrage
(appariement correct à gauche, choix d'une contre allée au milieu et à droite)

#### 7.5.2.4 Exemples d'incohérence entre les représentations

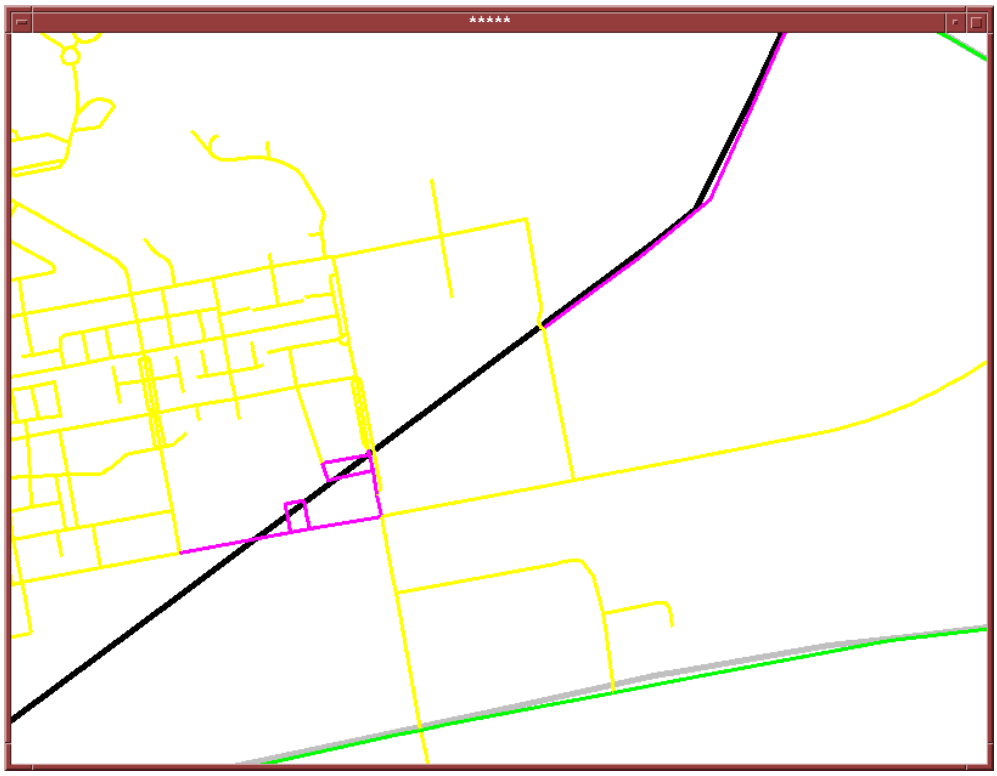


figure 98 : Exemple de tronçon BD CARTO manquant

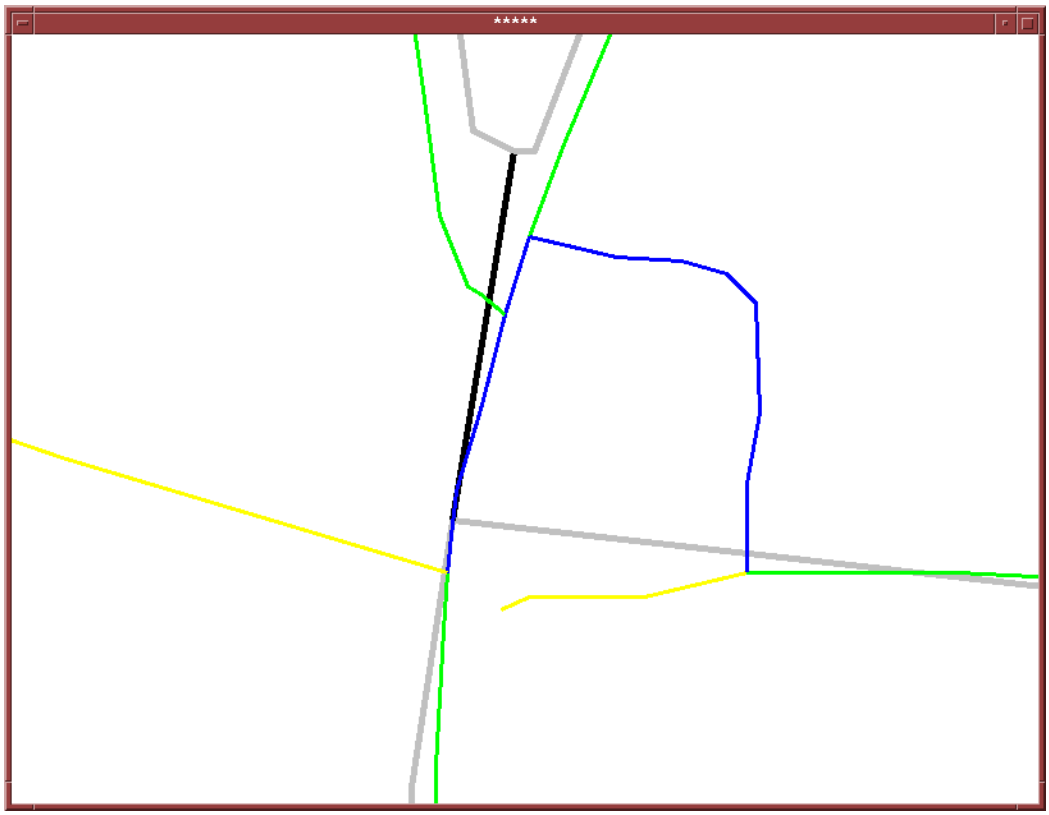



figure 99 : Exemple de discontinuité du réseau GEOROUTE entraînant
un appariement n-m détecté comme représentation incohérente

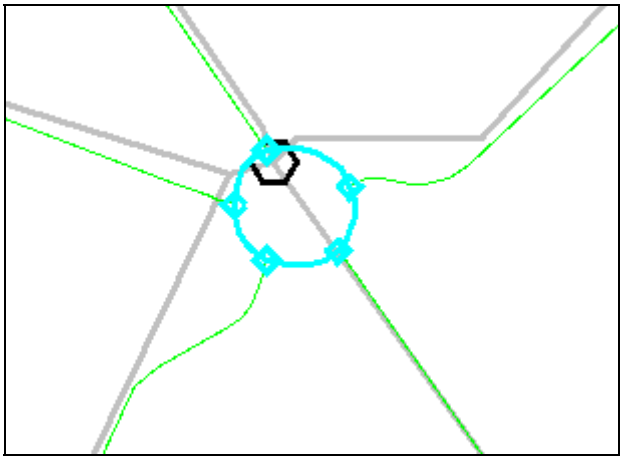


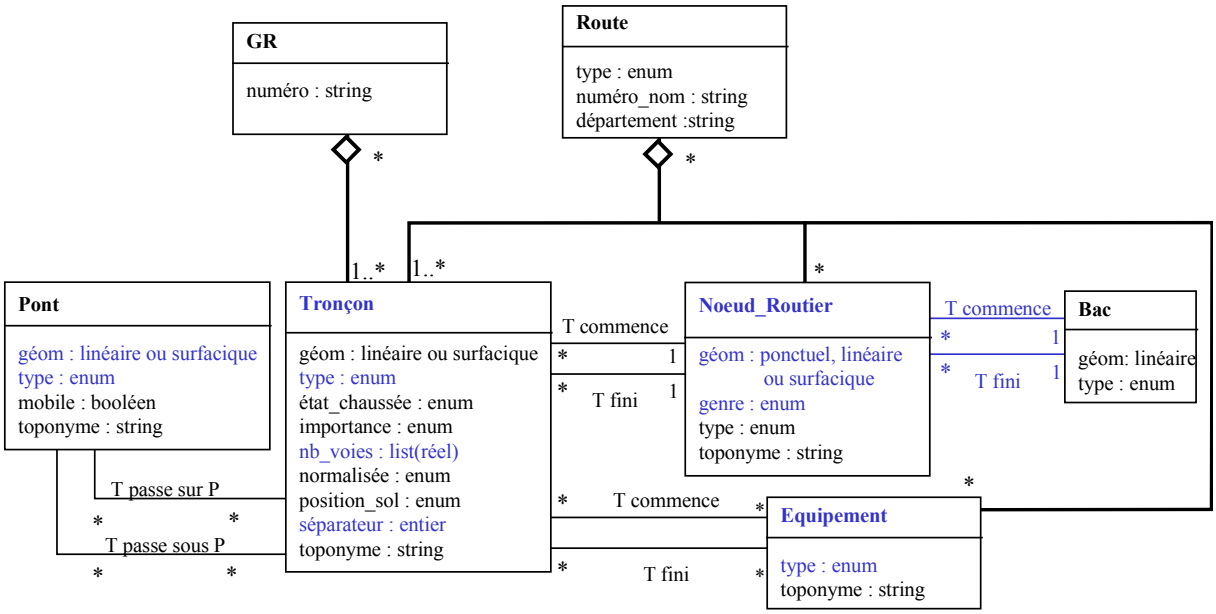
figure 100 : Exemple de défauts aux intersections de la BD CARTO
(tronçon parasite entre les deux noeuds)

## 7.6 Schémas des BDG de l'IGN

### *7.6.1 BD TOPO (après la pré-intégration)*

**pré-intégration**

fusion des classes Pont_Surf et pont,
fusion des classes Tronçon_route, Chemin, Sentier, Allée, Piste_Cyclable, Escalier, Escalier_Surf
fusion des objets de type chaussées séparés et du séparateur en un objet de type tronçon de route
fusion des classes Carrefour_na, Barrière, Péage
fusion des classes Parc_Stat, Aire
matérialisation des embarcadères




Tronçon = Tronçon_route, Chemin, Sentier, Allée, Piste_Cyclable, Escalier, Escalier_Surf

genre = nom classe d'origine

nb_voies = list (nb_voies sens direct, nb_voies sens indirect)
si il existe un séparateur et 2 tronçons partageant la même géométrie
list (nb_voies) sinon

séparateur = largeur du séparateur si il existe 0 sinon

Noeud_Routier = Carrefour_na, Barrière, Péage type = nom classe d'origine ou embarcadère

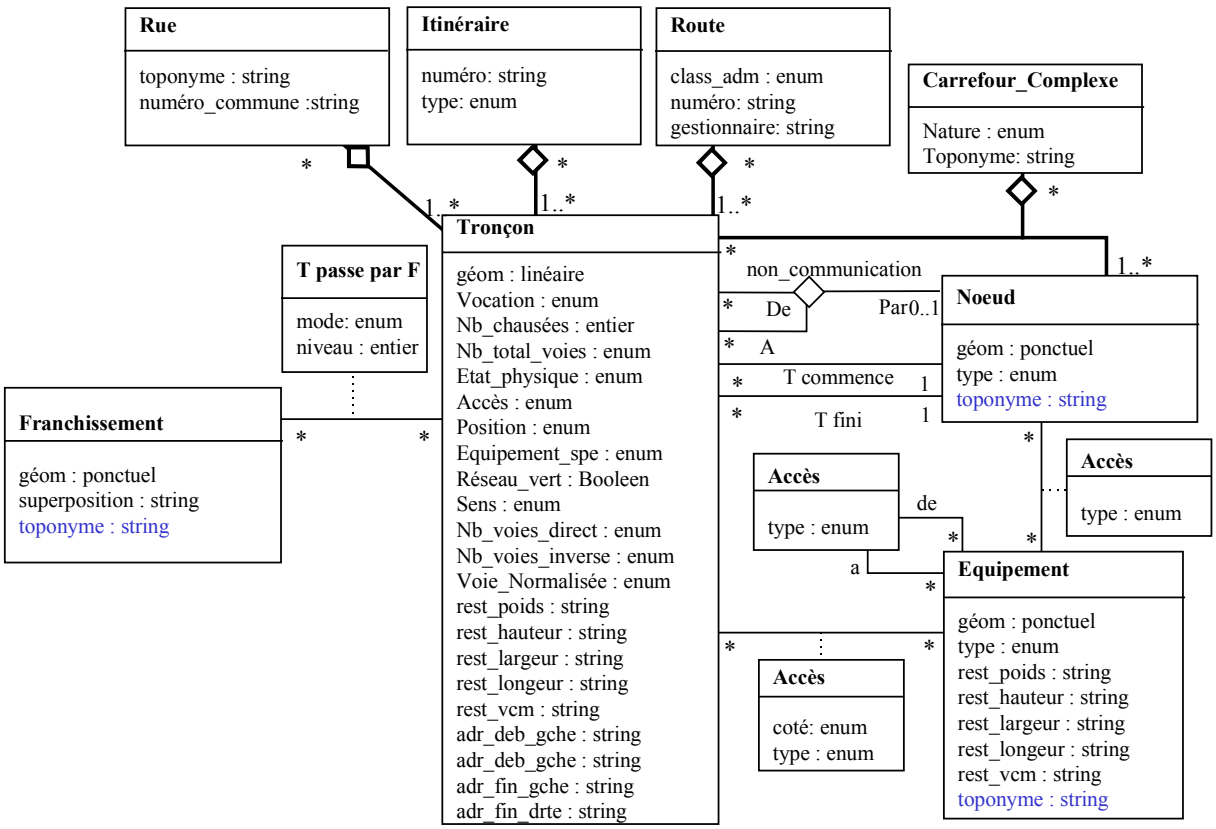
Equipement = Parc_Stat, Aire type = nom classe d'origine

Pont = Pont, Pont_Surf, Passerelle type = nom classe d'origine

### *7.6.2 Géoroute (après la pré-intégration)*

pré-intégration

fusion des classes de type toponyme avec les classes qu'elles nomment.




Pour la classe **Noeud,** le toponyme est issu de classe Topo_Noeud qui est relié à Noeud par une relation 1-1

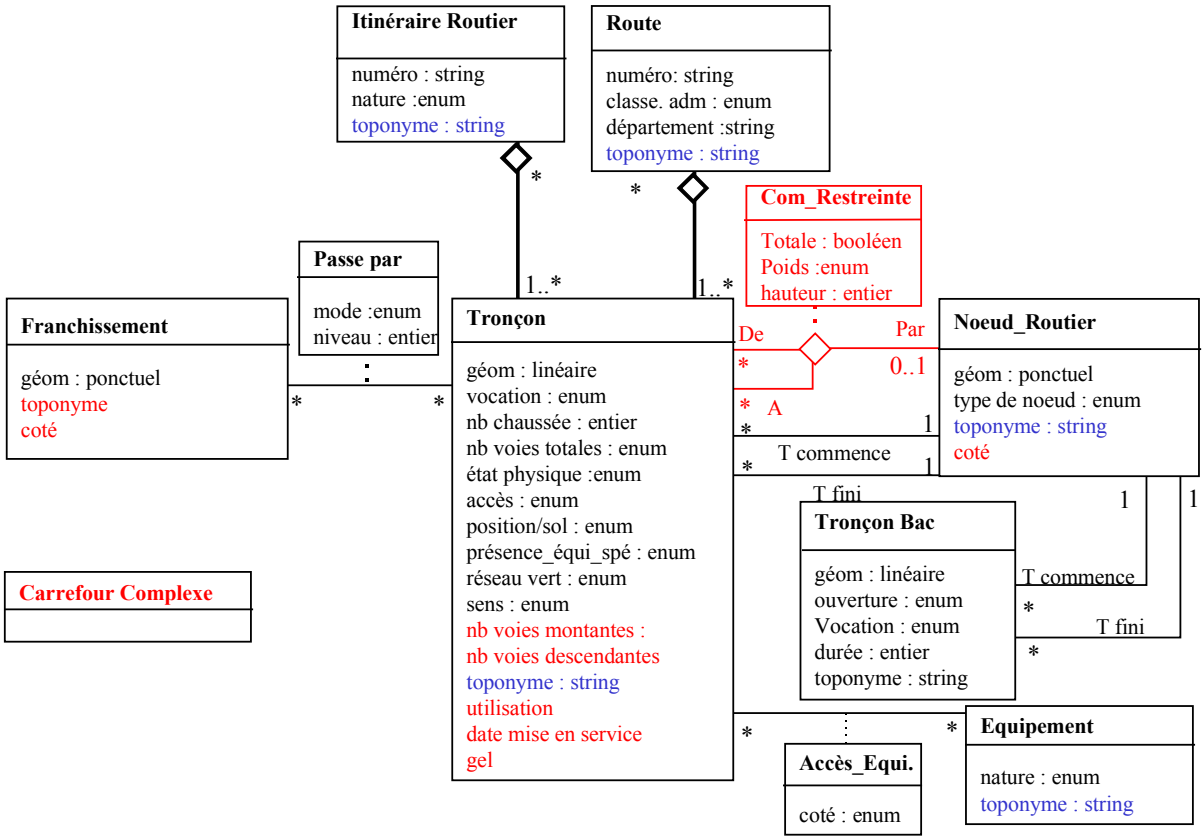
Pour la classe **Franchissement,** le toponyme est issu soit de la classe Topo_Franchissement, soit de la classe Topo_Tronçon si le tronçon portant le toponyme passe par un franchissement à un niveau > 0.

### *7.6.3 BD CARTO (après la pré-intégration)*

**pré-intégration**

ajout des franchissements déduits,

fusion des classes de type toponyme avec les classes qu'elles nomment.




En rouge : fait partie du schéma BD Carto V2 mais pas du schéma du jeu de données

## 7.7 Déclaration de correspondance entre les BDG de l'IGN

### *7.7.1 Déclaration des ACI entre la BD TOPO et GEOROUTE*

**Tronçon de route (TRONÇON)**

Remarques :

- La classe TRONÇON de la BD TOPO est issue de la fusion des classes tronçon_route dont les instances ont été normalisée (4.2.3) , Chemin, sentier, Allée, Piste_cyclable, Escalier, Escalier_surf.
- **BDT.type** = enum (« routier », « chemin », « sentier », « allée », « piste cyclable », « escalier »)
- **attributs virtuelle de GEOROUTE**
    - **G.vocation_V** = enum (« type autoroutier », « voie artérielle », « voie de distribution », « voie de desserte », « bretelle », « voie rapide urbaine »)
    - **G.etat_physique_V** = enum ( « route revêtue », « route non revêtue »), « en construction »)
    - **G.type_V** = enum («routier », « chemin ou sentier », « escalier »)

SET([[1:N] BDT.TRONÇON) ⊂ SET([[1:N] SELECTION ((¬ (Impasse(G.TRONÇON) ∧ L(G.TRONÇON)< 50 m)) ∨ (vocation ≠ « passerelle ») ) G.TRONÇON)

etat_chaussee = $_{L(G.Tronçon)>100m}$ etat_physique_V
(« revêtue », (« route revêtue »))
(« non revêtue », (« route non revêtue »))
(« en construction », (« en construction »))

importance CD $_{L(G.Tronçon)>100m}$ vocation_V
(« principale », (« type autoroutier », « voie rapide urbaine », « voie artérielle », « voie de distribution », « bretelle »))
(« secondaire », (« voie de distribution », « voie de desserte », « bretelle »))
(« quelconque », (« voie de desserte »))

nb_voies CD $_{card(BDT.Tronçon.nb\_voies) = 1 \wedge L(G.Tronçon)>100m}$ nb_total_voies
(1, (« 1 voie »),
(2, (« 2 voies », « 2 voies larges » ),
(3, (« 3 voies »))
(4, (« 4 voies »))
([5…∞], (« plus de 4 voies »))

nb_voies[0] CD $_{card(BDT.Tronçon.nb\_voies) = 2 \wedge Direction(BDT.Tronçon,G.Tronçon) \wedge L(G.Tronçon)>100m}$
nb_voies_sens_direct

nb_voies[0] CD $_{card(BDT.Tronçon.nb\_voies) = 2 \wedge \neg Direction(BDT.Tronçon,G.Tronçon) \wedge L(G.Tronçon)>100m}$
nb_voies_sens_inverse

nb_voies[1] CD $_{card(BDT.Tronçon.nb\_voies) = 2 \wedge Direction(BDT.Tronçon,G.Tronçon) \wedge L(G.Tronçon)>100m}$
nb_voies_sens_inverse

nb_voies[1] CD $_{card(BDT.Tronçon.nb\_voies) = 2 \wedge \neg Direction(BDT.Tronçon,G.Tronçon) \wedge L(G.Tronçon)>100m}$
nb_voies_sens_direct
(1, (« 1 voie »))
(2, (« 2 voies »))

(3, (« 3 voies »))
(4, ((« 4 voies »))
(5, ( (« 5 voies »)))
([6…∞],())

normalisee CF $_{L(G.Tronçon)>100m}$ voie_normalisee
(« normalisée », (« voie normalisée »))
(« non normalisée », (« voie non normalisée », « voie urbaine »))

type CF $_{L(G.Tronçon)>100m}$ type_V
(« routier », (« routier »))
(« chemin », (« chemin ou sentier »))
(« sentier », (« chemin ou sentier »))
(« escalier », (« escalier »))

position_sol CF position_sol
(« en surface », (« à niveau, normal », « sur viaduc, ou sur pont », « toboggan », « sur barrage », « sur digue »))
(« souterrain », (« en tunnel, souterrain »))

**Noeud routier (NOEUD) (voir 4.3.3.2)**

*Remarques* :

- La classe NOEUD de la BD TOPO est issue de la fusion des classes CARREFOUR_NA, PEAGE, BARRIERE, des embarcadères issues de la pré-intégration et des changements de commune issues de la pré-intégration . L'attribut énuméré « genre » à pour valeur le nom de la classe ou embarcadère ou encore changement de commune.
- **BDT.genre** = enum (« carrefour », « péage », « barrière », « embarcadère », « changement de commune »)
- **attribut virtuelle de la classe Noeud de la BD TOPO**
  - **Type_V** = enum (« carrefour ou cul de sac », « changement d'attribut », « péage », « barrière », « embarcadère », « changement de commune »)

BDT.Noeud ≡ $_{patte\_oie(G.SET) \wedge base\_patte\_oie(G.SET) < 50\ m}$ G.SET([3,3] Noeud, [3,3] Tronçon)
∧ BDT.SET([3,n] Noeud, [3,n] Tronçon) ≡$_{rond\text{-}point(BDT.SET) \wedge diamètre\_rond\text{-}point(BDT.SET) < 30\ m}$ G.Noeud
∧ SELECTION(BDT.Noeud.Type_V = « changement d'attribut ») BDT.Noeud ⊂ SELECTION(G.Noeud.Typenoeud = « changement d'attribut ») G.Noeud
∧ SELECTION(BDT.Noeud.Type_V = « péage ») BDT.Noeud ≡ SELECTION (G.Equipement.Nature = « péage »)Equipement
∧ SELECTION(BDT.NOEUD.Type_V ≠ « embarcadère ») BDT.Noeud ≡ $_{sinon}$ G.Noeud

BDT.Noeud.type_V CF G.Noeud.typenoeud
(« carrefour ou cul de sac », (« carrefour ou cul de sac », « intersection appartenant à rond-point », « intersection appartenant à carrefour aménagé », « intersection appartenant à carrefour aménagé avec tunnel et toboggan », « intersection appartenant à un échangeur complet »))
(« changement d'attribut », « changement d'attribut »)
(« barrière », (« barrière », « barrière de douane »)
(« changement de commune », « changement de commune »)

BDT.Noeud.Type_V CF G.Equipement.Nature = « péage »
(« péage »,(« péage »)

toponyme $_{\text{toponyme} \neq \varnothing}$ = toponyme

**Pont - Franchissement**

*Remarque* :
- La classe Pont de la BD TOPO est issue des classes PONT, PONT_SURF et PASSERRELLE, l'attribut énuméré type prend pour valeur le non de la classe d'origine.

SET ( [1,N] BDT.PONT) $\subseteq$ SET( [1,N] G.FRANCHISSEMENT) $\wedge$
SELECTION $_{(\text{type} = \text{« passerelle »})}$ BDT.PONT $\equiv$ SELECTION$_{(\text{vocation} = \text{« passerelle »})}$ G.TRONÇON

toponyme $_{\text{toponyme} \neq \varnothing}$ = toponyme

**Equipement - Equipement**

*Remarque* :
- La classe Equipement de la BD TOPO est issue des classes PARC_STAT et AIRE. , l'attribut énuméré Type prend pour valeur le non de la classe d'origine.

BDT.Equipement $\subset$ G.Equipement

Type CD Nature
(« parc_stat »,(« parking »)
(« aire », (« aire de service », « aire de péage »))

**Route - Route**

SELECTION (BDT.Route.type $\neq$ «nommee ») BDT.Route $\cap$ G.Route
AIC numero = numero
departement = gestionnaire
type CF classement_adm
(« autoroute », (« autoroute »))
(« nationale », (« nationale », « voie express »))
(« départementale », (« départementale », « voie express »))

SELECTION (BDT.Route.type = «nommee ») BDT.Route $\subset$
SELECTION(G.Itineraire.nature = « itineraire nommée ») G.Itineraire
AIC numero = numero

BDT.GR $\subset$ SELECTION(G.Itineraire.nature = « GR») G.Itineraire
AIC numero = numero

#### 7.7.1.1 Description des attributs compliqués à générer pour les classes de la BDI 1

**Tronçon**

**Type** = enum (« routier », « chemin », « sentier », « piste cyclable », « allée », « escalier »)

**Vocation** = enum (« type autoroutier », « voie artérielle », « voie de distribution », « voie de desserte », « bretelle », « voie rapide urbaine »)

**Etat_physique** = enum ( « route revêtue », « route non revêtue », « en construction »)

**Nb_voies =** tuple (sens direct : real, sens inverse : real)

le type des sous attribut est réel pour accepter la valeur une voie et demi.

**Normalisee** = enum (« voie normalisée », « voie non normalisée », « voie urbaine »)

**Position_sol** = enum (« à niveau, normal », « sur viaduc, ou sur pont », « toboggan », « sur barrage », « sur digue », « en tunnel, souterrain »)

**Noeud**

**Type** = enum (« carrefour ou cul de sac », « intersection appartenant à rond-point », « intersection appartenant à carrefour aménagé », « intersection appartenant à carrefour aménagé avec tunnel et toboggan », « intersection appartenant à un échangeur complet », « rond-point simple », « changement d'attribut », (« barrière », « barrière de douane », « changement de commune », « péage »)

## *7.7.2 Déclaration des ACI entre la BDI 1 (BDI) et la BD CARTO (BDC)*

**Tronçon de route (TRONÇON)**

Trois **classes virtuelles** sont crées pour la BD CARTO

- Vocation_V = enum (« autoroute », « grande circulation », « liaison régionale », « liaison locale », « bretelle »)
- Type_V = enum ( « routier », « piste cyclable », « chemin », « sentier »)
- Etat_Physique_V = (« route revêtue, route non revêtue, en construction »)

SET([[1:N] BDI.Tronçon) ⊃ BDC.Tronçon

nb_voies.sens_direct + nb_voies.sens_inverse
CD $_{L(BDI.TRONÇON) > 1000\ m \wedge nb\_chaussees = « 1\ chausée »}$ nb_total_voies
((0+1), (« 1 voie »))
((1+0), (« 1 voie »))
((1+1), (« 2 voies », « 2 voies larges »))
((2+0), (« 2 voies », « 2 voies larges »))
((0+2), (« 2 voies », « 2 voies larges »))
((2+1), (« 3 voies »))
((1+2), (« 3 voies »))
((1,5+1,5), (« 3 voies »))
((0+3), (« 3 voies »))
((3+0), (« 3 voies »))
((2+2), (« 4 voies »))
((1+3), (« 4 voies »))
((3+1), (« 4 voies »))
((0+4), (« 4 voies »))
((4+0), (« 4 voies »))
((a + b), (« plus de 4 voies »)) avec a+b ≥ 5

nb_voies.sens_direct CF $_{L(BDI.TRONÇON) > 1000\ m \wedge nb\_chaussees = « 2\ chausées » \wedge Direction(BDI.Tronçon,BDC.Tronçon)}$ nb_chaussee_montante

nb_voies.sens_direct CF $_{L(BDI.TRONÇON) > 1000\ m \wedge nb\_chaussees = « 2\ chausées » \wedge \neg Direction(BDI.Tronçon,BDC.Tronçon)}$ nb_chaussee_descendante

nb_voies.sens_inverse CF $_{L(BDI.TRONÇON) > 1000\ m \wedge nb\_chaussees = « 2\ chausées » \wedge Direction(BDI.Tronçon,BDC.Tronçon)}$ nb_chaussee_descendante

nb_voies.sens_inverse CF $_{L(BDI.TRONÇON) > 1000\ m \wedge nb\_chaussees = « 2\ chausées » \wedge \neg Direction(BDI.Tronçon,BDC.Tronçon)}$ nb_chaussee_montante
(1, (« 1 voie »))
(2, (« 2 voies »))
(3, (« 3 voies »))
(4, (« 4 voies »))
([5…∞], (« plus de 4 voies »))

Vocation CD Vocation_V
((« autoroute »,(« autoroute »)),
(« voie rapide urbaine », (« autoroute », « grande circulation »),
(« artérielle »,(« grande circulation »)),

(« distribution », (« grande circulation », « liaison locale »)),
(« desserte »,(« liaison locale »),
(« bretelle », (« bretelle »)),

Etat_Physique = $_{L(BDI.TRONÇON) > 1000\ m}$ Etat_Physique_V

Position_sol CD Position_sol
((« à niveau, normal », (« normal », (« en tunnel, souterrain, couvert ou semi couvert »))
(« sur viaduc, ou sur pont », (« sur viaduc, ou sur pont »))
(« toboggan », (« normal »))
(« sur barrage », (« normal »))
(« sur digue », (« normal »))
(« en tunnel, souterrain », (« en tunnel, souterrain, couvert ou semi couvert »))

Type = Type_V

**Noeud Routier**

BDI.SET([1,n] Noeud, [0,m] Tronçon) ⊃ BDC.Noeud ∨

BDI. NOEUD ≡SELECTION$_{(Nature = gare\ de\ péage)}$ BDC.Equipement

Type CF Type
(« carrefour ou cul de sac », (« carrefour ou cul de sac », « changement d'attribut »))
(« intersection appartenant à rond-point », (« rond-point > 100 m », « rond-point < 100 m »))
(« intersection appartenant à carrefour aménagé », (« carrefour aménagé > 100m »))
(« intersection appartenant à carrefour aménagé avec tunnel et toboggan », (« carrefour aménagé avec toboggan ou passage inférieur »))
(« intersection appartenant à échangeur complet », (« échangeur complet »))
(« changement d'attribut », « changement d'attribut »))
(« péage »,())
(« barrière », (« barrière »))
(« barrière de douane », (« barrière de douane »))
(« embarcadère », (« embarcadère », « embarcadère hors territoire »))
(« changement de commune »,()))

toponyme = toponyme

**Equipement**

SELECTION $_{(Type = « aire »)}$ Equipement ≡ SELECTION $_{(Nature \in \{« aire\ de\ repos »,« aire\ de\ service »\})}$ BDC.Equipement

Type CF Nature
(« aire », (« aire de repos », « aire de service »))

Toponyme = Toponyme

**Pont - Franchissement**

*Remarques* :

- Les instances de la classe franchissement de la BDC sont complétés par les instances implicites.
- Les toponymes de pont, viaduc, tunnel sont recherchés sur le tronçon de route associé passant sur le franchissement (méthode toponyme).

SET ( [1,N] BDI.PONT) ⊆ SET( [1,N] BDC.FRANCHISSEMENT)

Toponyme = Toponyme

**Route, GR, Itinéraire**

BDI.Route ⊂ BDC.Route
AIC numero = numero
departement = gestionnaire
BDI.Route ⊂ SELECTION(BDC.Itineraire_Routier ≠ « GR ») BDC.Itineraire_Routier
AIC numero = numero

BDI.GR ⊂ SELECTION(BDC.Itineraire_Routier = « GR ») BDC.Itineraire_Routier
AIC numero = numero

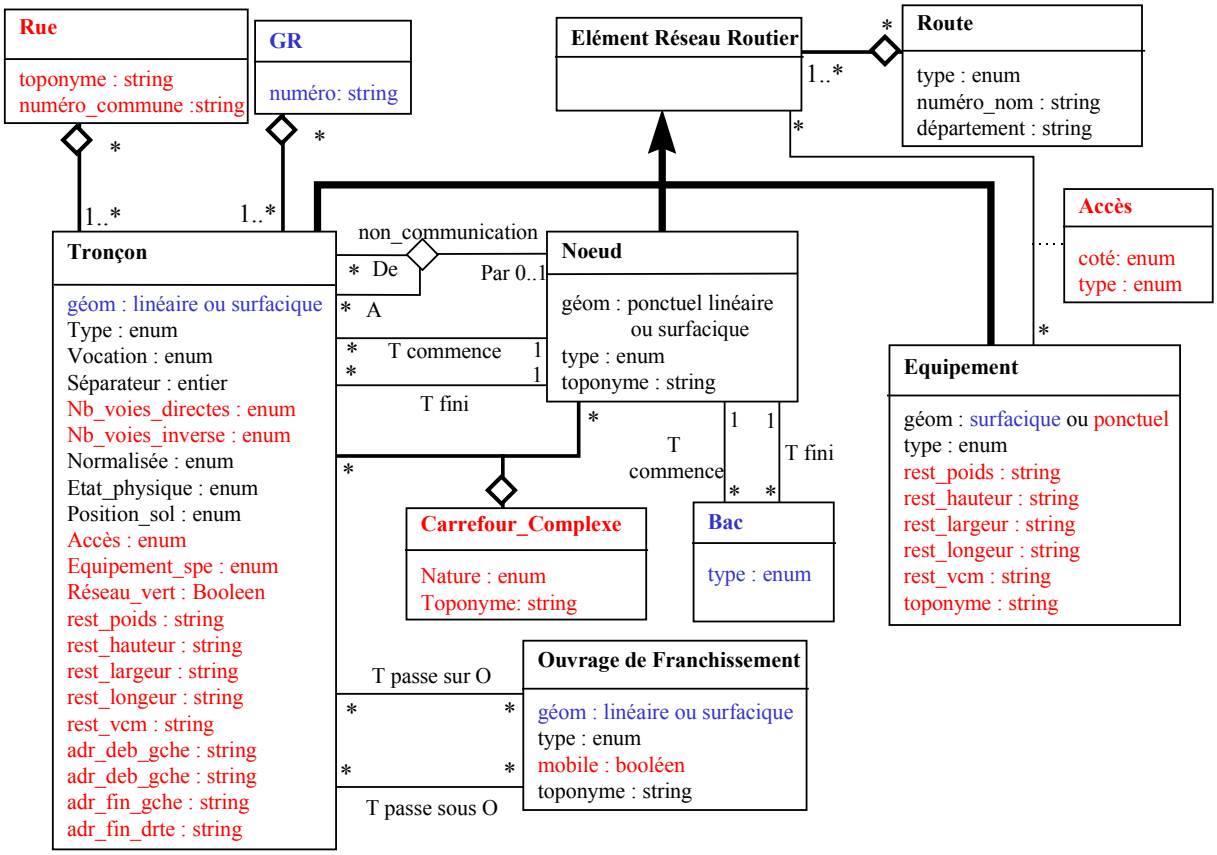
**Tronçon de Bac**

BDI.Bac ⊂ BDC.Tronçon_Bac

Type CF Vocation
(« autos », (« inconnu », « piétons et automobiles »)
(« piétons », (« inconnu », « piétons seulement »)

## 7.8 Schémas de la BD Intégrée

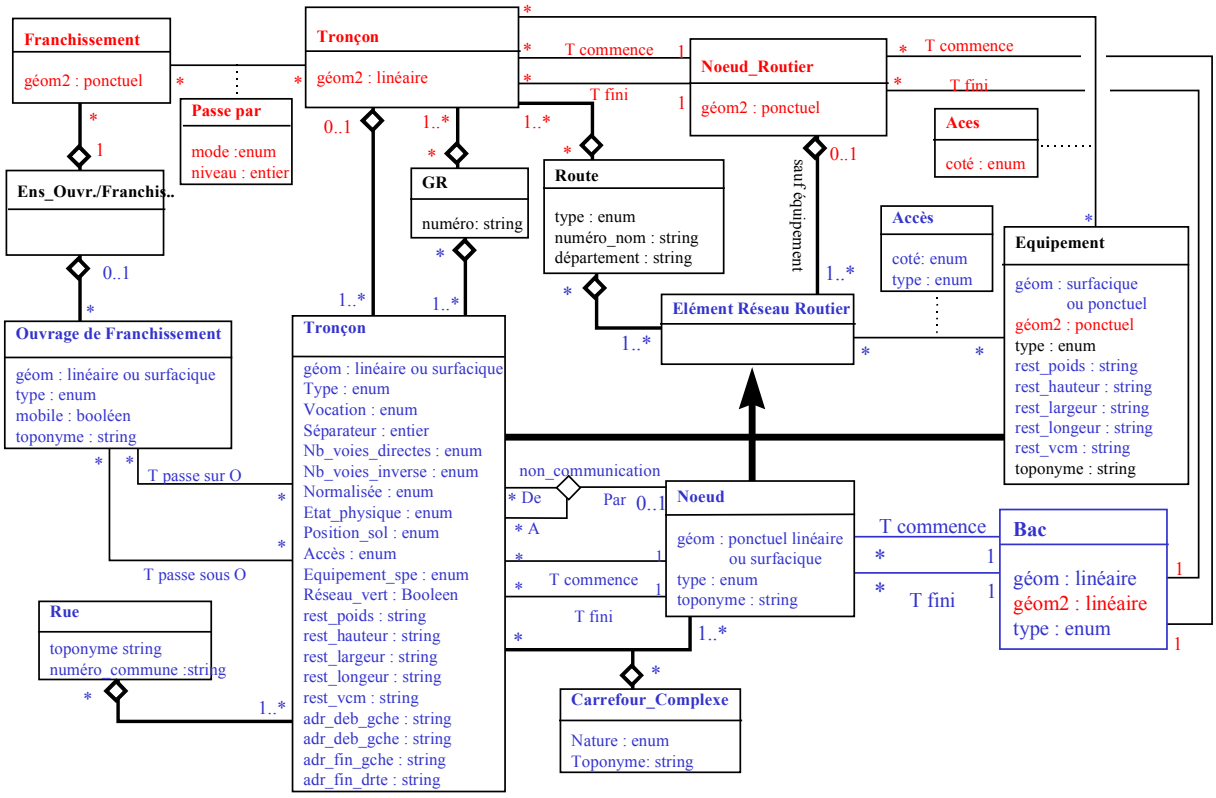
### *7.8.1 BD Intégrée 1*




en rouge pur GEOROUTE

en bleu pur BD TOPO

## 7.8.2 BD Intégrée 2



# 8. Bibliographie

**I. Abbas (1994)** Base de données vectorielles et erreur cartographique : problèmes posées par le contrôle ponctuel ; une méthode alternative fondée sur la distance de Hausdorff : le contrôle linéaire, Thèse de l'université de Paris 7.

**D. Abel, P. Kilby et M. Cameron (1994 a)** A Federates Systems Approach to Design of Spatial Decision Support System, Spatial Data Handling (SDH), Edinburgh, pages 46-59.

**D. Abel, P. Kilby et J. Davis (1994 b)** The systems integration problem, Int. Journal on Geographical Information Systems (IJGIS), Vol. 8, num. 1, pages 1-12.

**D. Abel, K. Yap, R. Acklaland, M. Cameron, D. Smith et G. Walker (1992)** Environmental Decision Support System Project : An Exploration of Alternative Architecture for Geographic Information Systems, Int. Journal on Geographical Information Systems (IJGIS), Vol. 6, num. 3, pages 193-204.

**G. Abrantes (1996)** The object-oriented approach to the specification of Geographic applications, Second Joint European Conference (JEC), Rumor, McMillan et Ottens (Eds.), IOS Press, Barcelonne, pages 48-257.

**J.-G. Affholder (1996)** Eléments de réflexion sur la terminologie des opérateurs de généralisation, rapport technique IGN.

**J. Alm, R. Lempien, A. Turkia et A. Talkakari (1994)** Database for supporting spatial data integration in SILMU research programme, 5th European Conference and Exhibition on Geographical Information Systems (EGIS) , Paris, pages 1186-1193 ou http ://www.aka.fi/silmu/tere.htm

**H. Alt, B. Behrends et J. Blomer (1992)** Approximate Matching of Polygonal Shapes, 7th Annual ACM Symposium. on Computational Geometry, pages 186-193.

**H. Alt et Gadau (1995)** Measuring the resemblance of polygonal curves. 8th Annu. ACM Symposium on Computational. Geometry, pages 102-109.

**R. Areia (1996)** Base de données multi-échelles, amélioration de l'appariement entre le réseau routier BD CARTO et BD Routière, DESS Application de l'informatique aux Sciences de la Terre, Paris VI.

**T. Atwood, J. Duhl, G. Ferran, M. Loomis et D. Wade (1993)** The object database standard : ODMG 93, R. G. G. Cattell (Ed.),Morgan Kaufmann, San Mateo, California.

**D. Baldwin, et A. Perera (1995)** Multi-scaled biodiversity management - Spatial Tools and Techniques, the Annual Meeting of the Society of American Foresters, http://www.spectranet.ca/~flep/flagpub.htmhttp://www.spectranet.ca/~flep/flagpub.htm

**L Band. (1994)** Multiple scale simulations of forest water and carbon flux in Ontario, the fall meetings of the American Geophysical Union.

**C. Batini, M. Lenzerini et S. Navathe (1986)** A comparative Analysis of methodologies for Database Schema Integration, ACM Computing Surveys, vol. 15, num. 4, pages 323-364.

**C. Bauzer Medeiros et G. Jomier (1993)** Managing alternatives and Data Evolution in GIS, ACM/ISCA Workshop on Advances in Geographic Information Systems.

**B. Bederson et J. Hollan (1994)** Pad ++: A zooming graphical interface for exploring alternate interface physics, ACM UIST'94, ACM Press, pages 17-26.

**A. Bel Hadj Ali (1997)** Etude de l'appariement des surfaces et implémentation des indicateurs de qualité en vue de leur sensibilité, Rapport de DEA SIG réalisé au COGIT, ENSG et Université Marne-la-Vallée.

**D. Bennett et D. Sharpe (1995)** Multi-scale Integrative Modeling for Ecosystem Management, Cache River Watershed, Illinois, GIS/LIS'95, pages 53-61.

**C. Berge (1983)** Graphes, Gauthier-Villars, 3ième édition.

**M. Bertolotto, L. De Floriani et E. Puppo (1994)** Multiresolution topological maps, Advanced Geographic Data Modelling (AGDM), Delft (Pays-Bas), pages 179-190.

**Y. Bonin (1995)** Contrôle de la précision géométrique de la BD CARTO par la méthode du contrôle linéaire, IGN/SDT/PNPV, num. QBDC-YB08.

**G. Booch (1991)** Object Oriented Design with applications, chapitre 4 : Classification, Benjamin/Cummings Publishing Company, Inc., pages 133-151.

**G. Booch, J. Rumbaugh, I. Jacobson (1997)** Unified Modeling Language User Guide, Addison-Wesley ou http://www.rational.com/uml/references/docset.html.

**B. Bouchon-Meunier (1994)** La logique floue, collection Que sais-je ?, PUF, Num. 2702.

**J. Boudon (1993)** Le syndrome de la confiture, MARI 93, Paris, pages 86-89.

**J. Bradley (1996)** Multi-Scale Maps, http://www.c3.lanl.gov/~cjhamil/Browse/main.html.

**P. Branly (1997)** Mise en place d'un processus d'appariement dans le cadre de l'intégration des bases de données BD Topo et Géoroute, mémoire ingénieur réalisé au COGIT, Ecole Supérieur des Géomètres et Topographes (ESGT).

**K. Brassel et R. Weibel (1988)** A review and conceptual framework of automated map generalization, Int. Journal on Geographical Information Systems (IJGIS), Vol. 2, Num. 3, pages 229-244.

**M. Breunig and A. Perkhoff (1992)** Data and System Integration for Geoscientific Data, Spatial Data Handling (SDH), Charleston, pages 272-280.

**P. Brooker (1995)** Development of a Multiresolutional Integrated Land Cover Database for Scotland and the Application of GIS in Predictive Spatial Modeling, ESRI User Conference Proceedings.

**B. Brugger (1994)** Spatial Theory for the integration of resolution-limited Data, Thèse de l'université du Maine (E.U.)

**B. Brugger (1995)** Theory for the Integration of Scale and Representation formats : Major Concepts and Practical Implications, COSIT Vienne, pages 297-310.

**B. Brugger, R. Barrera, A. Frank, K. Beard et M. Ehlers (1989)** Research Topic on Multiple Representations, Workshop on Multiple Representations Initiative 3, National Center for Geographic Information and Analysis (NCGIA), pages 53-67.

**O. Bucaille (1997)** Mise à jour d'une Base de Données Géographiques exemple de la BD Carto à l'Institut Géographique National, mémoire ingénieur CNAM réalisé au COGIT.

**K. Buehler et L. McKee (1996)** The OpenGIS Guide, rapport technique OGIS Project Technical Committee, http://www.ogis.org/guide/guide1.htm.

**B. Buttenfield (1991)** A rule for describing line feature geometry, Map Generalization : B. Buttenfield et R. McMaster (Eds.), Harlow Essex England : Longman Scientific, Chapitre 3, pages 150-171.

**B. de Cambray (1994)** Etude de la modélisation de la représentation de l'information spatiale 3D dans les bases de données géographiques, Thèse de doctorat de l'université Paris 6.

**B. de Cambray et TS. Yeh (1994)** A Multidimensional (2D, 2.5D and 3D) Geographical Data Model, Sixth International Conference on Management of Data, Mc Graw-Hill, Bangalore, Inde, pages 317-336.

**A. Car et A. Frank (1994)** Modelling a Hierarchy of Space Applied to Large Road Networks, International Workshop on Advanced Research in Geographic Information Systems (IGIS), Springer-Verlag, Ascona, pages 15-24.

**W. Cellary et G. Jomier (1990)** Consistency of Versions in Object-Oriented Databases; 16th Very Large DataBase (VLDB), Brisbane, Australie.

**CEN/TC 287 (1995)** Geographic Information - Data description : Metadata, CEN/TC 287/WG2.

**CEN/TC 287 (1996)** Geographic Information - Reference Model, European Prestandard, Final Draft prENV 12009.

**H. Chikh-Zaghar (1994)** Intégration multi-source de données géographiques par transformations correctives élastiques sous contraintes, 5th European Conference and Exhibition on Geographical Information Systems (EGIS), Paris, 1519-1529.

**N. Chrisman et M. Lester (1991)** A diagnostic test for error in categorical map, Auto Carto 10, ACSM/ASPR, Vol. 6, Baltimore, pages 330-348.

**C. Claramunt et M. Mainguenaud (1996)** A Spatial Data Model for Navigation Knowledge, Spatial Data Handling (SDH), Delft (Pays-Bas), Kraak et Molenaar (Eds.), pages 767-784.

**C. Crépeau et T. Huet (1993)** Le projet SIGMIP : vers un serveur régional d'information géographique en Midi-Pyrénées, Facettes de SIG, Mappemonde, Maison de la Géographie, Belin, vol. 4.

**J. Dangermond (1989)** The organizational impact of GIS technology, Arc News Summer 25-6.

**B. David (1991)** Modélisation, représentation et gestion d'information géographique, thèse de doctorat de l'université de Paris 6.

**B. David (1997)** Qualité d'une base de données géographiques, concepts et terminologie, Bulletin d'information de l'IGN numéro 67, http/www.ign.fr/QUALI/BI67/INDEX.HTML

**B. David, L. Raynal et G. Schorter (1993 a)** Evaluation of the OO approach for Geographical Applications, rapport technique Esprit Project (AMUSING) num. 6881.

**B. David, L. Raynal, G. Schorter et V. Mansart (1993 b)** Why objects in a geographical DBMS ?, Advances in Spatial Databases, LNCS 692, pages 264-276.

**B. David, L. Raynal, G. Schorter et V. Mansart (1993 c)** $GéO_2$ modélisation objet de données, Revue de géomatique, Hermès, pages 265-281.

**U. Dayal et H. Hwang (1984)** View definition and Generalization for Database Integration in Multidatabase System. IEEE Transactions on Software Engineering. SE10, pages 628-644.

**A. Demirkesen et B. Schaffrin (1996)** Map Conflation : Spatial point data merging and transformation, GIS/LIS'96, Denver, pages 393-404.

**J. Denègre et F. Salgé (1996)** Les Système d'information Géographique, Collection Que sais-je ?, Num. 3122, Presses Universitaires de France.

**T. Devogele (1997)** Les SIG Multi-représentations : Applications et Systèmes, Informatique pour l'Environnement Umweltinformatik'97 , Geiger, Jaeschke, Rentz, Simon Spengler, Zilliox Zundel (Eds.), Metropolis, Strasbourg, pages 678-687.

**T. Devogele, C. Parent et S. Spaccapietra (1998)** On Spatial Database Integration, Int. Journal on Geographical Information System (IJGIS), à paraître dans le vol. 12, Taylor & Francis.

**T. Devogele et L. Raynal (1995)** Intégration de bases de données à différentes échelles : Les liens de correspondances, Secondes Journées de la Recherche du GDR 1041 du CNRS, CASSINI, Jeansoulin (Ed.), Marseille.

**T. Devogele et L. Raynal (1996)** Modeling a Multi-Scale Database with Scale-Transaction Relationships, First International Conference on Geographic Information Systems in Urban Regional and Environmental Planning, SAMOS, Greece, T. Sellis (Ed.), pages 83-93.

**T. Devogele, J. Trevisan et L. Raynal (1996 b)** Building a multi-scale database with scale-transaction relationships, Spatial Data Handling (SDH), Delft (Pays-Bas), Kraak et Molenaar (Eds.), pages 6.19-6.33.

**T. Devogele, J. Trevisan et L. Raynal (1996 c)** Processus de constitution d'une base de données multi-échelles, Revue Internationale de géomatique, Vol. 6, Num. 2-3, pages 249-263.

**E. Dijkstra (1959)** A note on Two Problems in Connection with Graphs. Numer Math., Vol. 1, pages 269-271.

**Direction technique de l'IGN (1991)** Normalisation des Références de coordonnées (RIG), note DT /514 du 4 juillet.

**R. Djadri (1996)** Diagrammes de Voronoï généralisés, rapport de stage DEA IFA au COGIT, université de Marne la Vallée, réalisé au COGIT.

**J. Dougenik (1980)** WHIRLPOOL : A geometric processor for polygon coverage data, Auto Carto 4, pages 304-311.

**P. Dumolard, L. Raynal, Y. Demazeau et J.-P. Lagrange (1995)** Axe B : S.I.G. Multi-échelle, http://lieu.univ-mrs.fr/GDR-CASSINI/axeB.htm#rapport95.

**Y. Dupont (1994)** Resolving Fragmentation Conflicts in Schema Integration, 13th Int. Conf. on The Entity Relationship Approach.

**Y. Dupont (1995 a)** Problématique et résolution contextuelle des conflits de fragmentation dans l'intégration de schémas Revue Ingénierie des Systèmes d'Information, numéro spécial Bases de Données Avancées, Vol. 3, Num. 1, 1995.

**Y. Dupont (1995 b)** Une méthode flexible pour l'intégration de schémas dans les bases de données à objets complexes, thèse de Ecole Polytechnique Fédérale de Lausanne, Lausanne, Suisse.

**Y. Dupont (1996)** Une méthode flexible pour l'intégration de schémas dans les bases de données à objets complexes, BDA'96, Cassis.

**M. Egenhofer, A. Frank et J. Jackson (1989)** A topological data model for spatial databases, Symposium Design an Implementation of Large Spatial Databases, Springer-Verlag, pages 271-286.

**T. Eiter et H Mannila (1994)** Computing Discrete Fréchet Distance, rapport technique Christian Doppler Labor für Expertensensyteme, Université Technique de Vienne, num. CD-TR 94/64.

**Equipe BD CARTO (1994)** Spécification de contenu BD CARTO rapport technique de l'IGN.

**Equipe BD TOPO (1994)** La BD Topographiques : Spécifications détaillées, version 3.1, rapport technique IGN / SIT, num. SIT/2320.

**ESRI France (1997)** GeoKiosk la borne interactive vue par ESRI France, La lettre trimestrielle d'ESRI France num. 11, pages 5-6.

**J. Euzenat (1994)** Granularité dans les représentations spatio-temporelles, rapport techniques de l'INRIA, num. 2242.

**J. Euzenat (1995)** An algebraic approach for granularity in qualitative time representation, 14th International Joint Conference on Artificial Intelligence, Montréal (CA), pages 894-900.

**G. Fagan et H. Soehngen (1987)** Improvement of GBF/DIME File Coordinates in a Geobased Information System by Various Transformation Methods and Rubbersheeting Based on Triangulation, Auto Carto 8, Baltimore, pages 481-491.

**Federal Geographic Data Committee (1994)** Content standards for spatial Metadata.

**P. Ferreira (1996)** Intégration de schémas de bases de données géographiques orientés objet à différentes échelles, rapport de DEA SIG réalisé au COGIT, IGN/ENSG.

**R. Flowerdew (1992)** Geographic Information Systems, chapitre 24 : Spatial Data Integration, Maguire, Goodchild, Rhind (Ed.), Longman Scientific & Technical, pages 337-358.

**R. Flowerdew et S. Openshaw (1987)** A review of the problems of transferring data from one set of areal units to another incompatible set, rapport de recherche nun. 4, Northern RRL.

**A. Frank (1987)** Overlay processing in spatial information systems, Auto Carto 8, Baltimore.

**A. Frank (1990)** Acquiring a digital base map - A theoretical investigation into the problems of sharing data, Rapport technique université du Maine.

**A. Frank (1991)** Advances in cartography, chapitre Design of cartographic databases, Elsevier applied science, International cartographic association, pages 15-44.

**A Frank et S. Timpf (1994)** Multiple representations for cartographic objects in a multi-scale tree an intelligent graphical zoom, Comput. & Graphics, Vol. 18, Num. 6, pages 823-829.

**S. Fraser (1995)** Multi-Scale Analysis of Forest Landscape, rapport technique University, Burnaby, B. C, V5A 1S6, http://www.res.for.gov.bc.ca/news/frbc/proposal/ndp_1.html.

**M. Fréchet (1906)** Sur quelques points du calcul fonctionnel, Rendiconti del Circolo Mathematico di Palermo, 22 :1-74.

**J. Frew, L. Carver, C. Fischer, M. Goodchild, M. Larsgaard, T. Smith et Q. Zheng (1995)** The Alexandria Rapid Prototype: building a digital library for spatial information, ESRI User Conference, ESRI.

**G. Furmas et B. Bederson (1995)** Space-Scale Diagrams : Understanding Multiscale Interfaces, http://community.bellcore.com/gwf/CHI95.htm.paper/gwf_body_thumb.html.

**Y. Gabay and Y. Doytsher (1994)** Automatic adjustment of line maps, GIS/LIS, lieu, pages 233-241.

**O. Gascuel (1987)** Plage : Un outil pour construire des systèmes d'apprentissage. Intellectica, 1(2/3).

**A. Gagnon et J. Malboeuf (1994)** Dictionnaire des entités géographiques au gouvernement du Québec, 2ième édition, Ministère des ressources naturelles, Gouvernement du Québec.

**Geller, Y. Perl, E. Neuhold et A. Seth (1992)** Structural Schema Integration With Full and Partial Correspondence using the Dual Model, Information Systems, Vol. 17, Num. 6.

**GEO2DIS (1997)** GEO2DIS : a Client-Server architecture, on Internet, to document and access geodata stored on heterogeneous GISs. http://www.pisa.intecs.it/projects/GEO2DIS/.

**GEOROUTE (1994)** format de livraison GEOROUTE, version A. version1.3, IGN.

**GeoSystems Global Corp. (1996)** Interactive Atlas Map Quest, http://www.mapquest.com/.

**D. Gillman (1985)** Triangulation for Rubber-sheeting, Auto Carto 7, Washington, pages 191-199.

**GIS/Trans Ltd (1995)** GIS Conflation and Its Automation, rapport technique GIS/Trans, Ltd, Cambridge, MA.

**C. Gold (1990)** Space revisited - back to the basics, Spatial Data Handling, Zurich, pages 175-189.

**M. Goodchild (1990)** Geographical Data Modeling, Two Perspectives on Geographic Data Modeling, rapport technique 90-11 du National Center for Geographic Information and Analysis, Santa Barbara (NCGIA), CA.

**M. Goodchild (1991)** Issue of quality and uncertainty, Advances in Cartography, Müller Ed., Barking, Essex/ Elsevier, pages 113-139.

**W. Gotthard, P. Lockemann et A. Neufeld (1992)** A System-Guided View Integration for Object Oriented Databases, IEEE Transactions on Knowledge, Vol. 4, Num. 1, pages 1-22.

**Gouvernement du Québec (1992)** Prototype sur l'intégration et l'échange de données à référence spatiale, rapport technique du Ministère de l'énergie et des ressources naturelles du Québec.

**Gouvernement du Québec (1996)** Actualisation des données à référence spatiales (volets échange et intégration), rapport technique du Ministère de l'énergie et des ressources naturelles du Québec.

**C. Grateau (1994)** Contrôle qualité d'une base de données géographique, Rapport technique IGN/SIT numéro 0636.

**S. Grumbach, I. Herlin et E. Simon (1996)** Comment faire entrer la géographie dans une base de données, rapport annuel de l'INRIA 1995.

**O. Günter (1989)** Database Support for Multiple Representations, Workshop on Multiple Representations Initiative 3, National Center for Geographic Information and Analysis (NCGIA), pages 50-51.

**J.-F. Hangouët (1995)** Computation of the Hausdorff distance between plane vector polylines, Auto Carto 12, Charlotte, pages 1-10.

**F. Harvey et F. Vauglin (1996)** Geometric Match-processing : Applying Multiple Tolerances, Spatial Data Handling (SDH), Delft (Pays-Bas), Kraak et Molenaar (Eds.), pages 155-171.

**F. Hausdorff (1919)** Dimension und ausseres, Mass. Mathematische Annalen, num. 79, pages 157-179.

**S. Hayne et S. Ram (1990)** Multi-user view integration system (muvis) : An expert system for view integration, Data Engineering, IEEE, pages 402-409.

**G. Higgs, P. Longley et D. Martin (1994)** A Multi-Scale Analysis of the Impacts of the Council Tax : A GIS Approach, rapport technique Ulster Business School, http://www.ulst.ac.uk/faculty/UBS/sect19.txt.

**IGN (1996 a)** La BD Cartographiques : Descriptif technique, rapport de la direction commerciale de l'IGN, 136 bis, rue de Grenelle, 75 700 Paris 07.

**IGN (1996 b)** La BD Topographiques : Descriptif technique, rapport de la direction commerciale de l'IGN, 136 bis, rue de Grenelle, 75 700 Paris 07.

**IGN (1996 c)** Géoroute : Descriptif technique, rapport de la direction commerciale de l'IGN, 136 bis, rue de Grenelle, 75 700 Paris 07.

**ISO (1994)** Industrial automation systems and integration - Product data representation and exchange Part 11 : Description methods : The EXPRESS language reference manual, rapport technique num. 10303-11:1994(E).

**ISO/TC 211/WG 1 (1996)** Conceptual Schema Language rapport technique num. ISO 15046-3, Working Draft.

**O. Jamet et B. Phalakarn (1989)** Un algorithme rapide de mise en correspondance de cartes, AFCET, 7$^{eme}$ congrès de reconnaissance des formes et intelligence artificielle, Barcelone, pages 1123-1130.

**D. Jardine et S. Yazid (1989)** Integration of Information Submodels, Information system concepts : An In depth Analysis, IFIP, Falkenberg, Lindgreen, Elsevier (Ed.), Science Publishers B.V, North Holland, pages 247-267.

**C. Jones. (1991)** Database architecture for multi-scale GIS, Auto Carto 10, volume 6, ACSM/ASPRS, Baltimore, pages 1-14.

**C. Jones, D. Kidner, L.Q. Luo, G. Bundy et J.M Ware (1996)** Database design for a multi-scale spatial information system, Int. Journal on Geographical Information Systems (IJGIS) , Vol. 10, Num. 8, pages 901-920.

**C Jones, D. Kidner et J. Ware (1995)** Equivalence tests for integration of multi-source spatial data, Eurocarto 13, European Communities, Luxembourg, pages 13-22.

**C Jones, C. Taylor, D. Tudhope et P. Beynon-Davies (1996)** Conceptual, Spatial and Temporal Referencing of Multimedia Objects, Spatial Data Handling (SDH), Delft (Pays-Bas), Kraak et Molenaar (Eds.), pages 2.13-2.26.

**C. Karra, R. Haushild, A. leFeuvre et P. de Fraipont (1994)** Qualité écologique d'un milieu alluvial : essai de modélisation à l'aide d'un SIG sur la forêt Rhénane, 5$^{th}$ European Conference and Exhibition on Geographical Information Systems (EGIS), Paris, pages 1594-1603.

**M. Kavouras, D. Paradissis, R. Ecker et J. Jansa (1995)** Geographic Information Systems : Materials for a Post Graduate Course vol. 1 : Spatial Information, chapitre 5, Data Sources for GIS, A. Frank (Ed.), Technical University Vienna, vol. 4, pages 177-250.

**H. Kemppainen (1994)** Modelling Update propagation in spatial Databases, Int. Archives of Photogrammetry and Remote Sensing, Vol. 30, Num. 4, pages 625-632.

**H. Kemppainen et J. Albrecht (1996)** Scalability of spatial operators, GIS/LIS'96, American Society for Photogrammetry and Remote Sensing, Denver, pages 162-173.

**W. Kent (1981)** Data Model Theory Meets a Practical Applications, Very Large DataBase (VLDB), Cannes, France, pages 13-22.

**D. Kidner (1996)** Geometric Signatures for determining polygon equivalence during multi-scale GIS Update, Second Joint European Conference, IOS Press, pages 238-247.

**D. Kidner et C. Jones (1994)** A Deductive Object-Oriented GIS for handling multiple representations, Spatial Data Handling, pages 882-900.

**T. Kilpeläinen (1995)** Requirements of a Multiple representation database for topographical data with emphasis on incremental generalization, 17th ICA/ACI Barcelone.

**W. Kim, I. Choi, S. Gala et Scheevel (1993)** On Resolving Schematic Heterogeneity in Multidatabase Systems, Distributed and Parallel Databases, Vol. 1, Num. 3, pages 251-279.

**W. Kuhn (1994)** Defining semantics for spatial data transfers, Spatial Data Handling (SDH), Edinburgh, pages 973-987.

**J-P. Lagrange et A. Ruas (1994)** Geographic information modelling : GIS and generalisation, Spatial Data Handling (SDH), Edinburgh, pages 1099-1117.

**P. Lahti (1997)** Geographic Information Systems (GIS) as an integrative platform for economical management, Ingénierie des systèmes d'informations, vol. 5, num. 2, pages 241-252.

**B. Langou et M. Mainguenaud (1994)** Graph data model operations for network facilities in a geographical information system, Spatial Data Handling (SDH), Edinburgh, pages 1002 - 1019.

**J. Larson, S. Navathe, et R. Elmasri (1989)** A Theory of Attribute Equivalence in Databases with Application to Schema Integration, IEEE Transaction on Software Engineering, vol.15, num. 4, pages 449-463.

**R. Laurini (1995)** Programme de recherche en bases de données géographiques, Revue internationale de géomatique, Hermès, vol. 5, num. 1, pages 83-99.

**R. Laurini (1996)** Raccordement géométrique de bases de données géographiques fédérées, Ingénierie des systèmes d'informations, vol.4, num. 3, pages 361-388.

**R. Laurini et D. Thompson (1992)** Fundamentals of spatial information systems, The Apic Series, num. 37, Academic press.

**C. Lemarié (1996)** Etat de l'art sur l'appariement, rapport technique IGN, num. DT/9600022.

**C. Lemarié et L. Raynal (1996)** Geographic Data Matching : First Investigation for a Generic Tool, GIS/LIS, Denver, pages 405-420.

**H. Le Men (1996)** Généralisation cartographique pour l'occupation du sol : application au passage V1 à V2 pour la BD Carto®, Revue internationale de géomatique, Hermès, vol. 6, num. 2-3, page 227-248.

**H. Le Men et O. Jamet (1990)** Interprétation automatique de l'occupation du sol sur image SPOT, Symposium International de Cartographie Thématique Dérivée des Images Satellitaires, Saint Mandé.

**W. Litwin, L. Mark et N. Roussopoulos (1990)** Interoperability of multiple automous databases, ACM Computer Surveys, Vol. 23, Num. 3, pages 267-293.

**A. Lupien et W. Moreland (1987)** A general approach to map conflation, Auto Carto 8, Baltimore, pages 630-639.

**M. Lynch et A. Salford (1985)** Conflation : Automated map compilation. A video game approach, Auto Carto 7, Washington D.C, pages 343-352.

**D. Maguire, G. Stickler et G. Browning (1992)** Handling Complex Objects in Geo-relational GIS, Spatial Data Handling.

**M. Mannino et W. Effelsberg (1984)** Matching Techniques in Global Schema Design, IEEE international Conference on Data Engineering, Los Angeles, pages 418-425.

**M. Mannino, S. Navathe et W. Effelsberg (1988)** A Rule-Based Approach for Merging Generalization Hierarchies, Information Systems, Vol. 13, Num. 3, 1988.

**D. Mark (1989)** Multiple views of multiple representations, Workshop on Multiple Representations, National Center for Geographic Information and Analysis (NCGIA), Initiative 3, num. 89-3, pages 68-71.

**R. McMaster (1986)** A statistical Analysis of Mathematical Measures for Linear Simplification, The American Cartographer, vol. 23.

**R. McMaster and S. Shea (1992)**, Generalization in Digital Cartography, Association of American Geographers.

**MEGRIN (1996)** Seamless Administrative Boundaries of Europe (SABE) Technical Overview, http ://www.ign.fr/megrin/sabe/sabedesc.exe

**G. Michel (1996)** Expression des besoins et échanges de données géographiques sur le terrain autour du système VESPUCCI - Rapport de DEA SIG, IGN/ENSG réalisé à la CEGN.

**A. Montanari, E. Main, Ciapessoni et E. Ratto (1992)** Dealing with time and granularity in the event calculus, 4th FGCS, Tokyo, pages 702-712.

**B. Morel (1996)** Les données sur l'environnement et les ressources naturelles, les entretiens de Ségur, Les inforoutes de l'environnement, ministère de l'environnement, pages 14-16.

**A. Motro (1987)** Superviews; Virtual Integration of Multiple Databases, IEEE Transactions on Software Engineering 13 (7) pages 785-798.

**J.C. Müller (1991)** Generalization of spatial databases, Geographical information Systems Principles and Applications, Maguire, Goodchild and Rhind (Eds.), Publisher Longman Scientific & Technical, pages 457-475

**J-C Müller, J-P. Lagrange, R. Weibel et F. Salgé (1995)** Generalisation : State of the art and issues, GIS and GENERALISATION, GISDATA, Müller, Lagrange Weibel (Eds.) Taylor & Francis, pages 3-17.

**S. Mustière (1995)** Mesures de la qualité de la généralisation du linéaire. Rapport de stage du DESS de Cartographie, Université Paris I, ENSG réalisé au COGIT.

**R. Myers (1997)** Balancing the use of Technology and Field Study in Ecosystem Research : A Multi-Scale Approach, the Journal of the Washington Academy of Sciences, ou http://ice.ucdavis.edu/~robyn/poster.html.

**P. Noonan, et D. Cowen (1995)** Multi Scale Analysis of Property Sales Ratio Data, GIS/LIS'95,pages 788-794.

**Norme ISO 8402 (1994)** Management de la qualité et assurance de la qualité.

**T. Nyerges (1989)** Schema integration analysis for the development of GIS databases, Int. J. Geographical Information Systems, vol.3, num. 2, pages 153-183.

**$O_2$ (1991)** The $O_2$ System; in Communications of the ACM, vol. 34, num. 10.

**Open GIS Consortium (1996)** The OpenGIS Abstract Specification : An Object Model for Interoperable Geoprocessing, Revision 1, rapport technique Open GIS Consortium, num. 96-015R1.

**T. Öszu et P. Valduriez (1989)** Principles of distributed Database Systems. Prentice Hall, Englewood Cliffs.

**P. van Oosterom et V. Schenkelaars (1991)**.The development of an interactive multi-scale GIS, Int. Journal on Geographical Information Systems (IJGIS) , vol. 9, num. 5, pages 489-507.

**P. van Oosterom (1995)** The GAP-tree, an approach to 'on-the-fly' map generalisation of an area partitioning, GIS and GENERALISATION, GISDATA, Müller, Lagrange, Weibel (Ed.) Taylor & Francis, pages 120-132.

**A. Ouksel et C. Naiman (1993)** Cooperatrion in Heterogeneous Database System, Int. Workshop on next generation information technologies and System, Haifa, Israel,

**O. Ousset (1997)** Contrôle de cohérence et migration de données géographiques dans une base de données intégrée, Rapport de stage de DESS Applications Informatique aux Sciences de la Terre de l'université Pierre et Marie Curie, réalisé au COGIT.

**C. Parent et S. Spaccapietra (1996)** Intégration de bases de données : panorama des problèmes et des approches, Ingénierie des systèmes d'information, vol. 4, num. 3, pages 333-358.

**C. Parent et S. Spaccapietra (1997)** Database integration; an overview of issues and approaches, CACM, à paraître.

**C. Parent, S. Spaccapietra et T. Devogele (1996)** Conflicts in Spatial Database integration, Parallel and Distributed Computing Systems (PDCS), Dijon.

**L. Peltz et T. Liebermann (1995).**Comparison of Digital Hydrographic-Basin Boundaries at Several Scales, ESRI User Conference Proceedings.

**W. Peng, K. Tempfi et M. Molenaar (1996)** Automated Generalization in a GIS, Geoinformatic'96, West Palm Beach pages 135-146.

**A. Perera, R. MacArthur et D. Baldwin (1994)** A multi-scale evaluation of the protected area network of the Great Lakes forest region, the International Forest Biodiversity Conference.

**J. Perkal (1956)** On Epsilon length, Bulletin de l'académie Polonaise des Sciences. vol. 4, pages 399-403.

**Y. Perl, J. Geller, E.J. Neuhold et V. Turau (1989)** The Dual Model for Object-Oriented Databases, rapport CIS-91-30, New Jersey Institute for technology.

**D Peuquet (1981 a)** An examination of techniques for reformatting digital cartographic data / Part 1 : The raster to vector process, Cartographica, vol. 18, pages 34-48

**D. Peuquet (1981 b)** An examination of techniques for reformatting digital cartographic data / Part 2 : The vector to raster process, Cartographica, vol. 18, pages 21-33

**B. Phalakarn (1991)** Evaluation de la qualité des processus de segmentation d'image par mise en correspondance à une référence, thèse de doctorat de l'université de Paris 7.

**M. Piron (1992)** Analyse statistique d'un système d'échelles, rapport technique ORSTOM Paris VI.

**M. Piron (1993)** Changer d'échelle, une méthode pour l'analyse des systèmes d'échelles, L'espace géographique, vol.2, pages 147-165.

**J. Piwowar, E. LeDrew et D. Dudycha (1990)** Integration of spatial data in vector and raster formats in a geographic information system environment, Int. Journal on Geographical Information Systems (IJGIS), vol.4, num. 4, pages 429-444.

**C. Plazanet (1996)** Enrichissement des bases de données géographiques : analyse de la géométrie des objets linéaires pour la généralisation cartographique (application aux routes), Thèse de l'université de Marne-la-Vallée, Laboratoire COGIT.

**D. Preux (1995)** Réseau national des données sur l'eau, intégration de sources de données, Base de données et systèmes d'informations

**Projet BD Routière (1992)** Manuel de saisie, Version 2.4, rapport technique num. BDR-P16.

**D. Pullar (1993)** Consequences of using a tolerance paradigm in spatial overlay. Auto Carto 11, ACSM/ASPR, Minneapolis, pages 288-296.

**E. Puppo et G. Dettori (1995)** Towards a Formal Model for Multiresolution Spatial Maps, SSD'95, Egenhofer and Herring (Eds.), Portland, pages 153-169.

**J. Querzola et M. Billout (1995)** MERCATOR 1995, guide de la cartographie, EURO-Vista.

**R. Ramirez (1997)** Development of a common Framework to express raster and vector datasets, Auto Carto 13, ACSM/ASPRS, Seattle, pages 155-163

**L Raynal, B. David et G. Schorter (1995)** Building an OOGIS prototype : Experiments with $GéO_2$, Auto Carto 12, Charlotte, ACSM/ASPRS, pages 137-146.

**L. Raynal, P. Dumolard, G d'Aubigny, P. Rigaux, C. Scholl et D. Larcena (1996)** Gérer et générer des données spatiales hiérarchisées, Revue internationale de géomatique, Hermès, vol 6., num. 4, pages 365-382.

**L. Raynal et N. Stricher (1994)** Base de données multi-échelles : Association géométrique des tronçons de route de la BD CARTO et de la BD TOPO, 5th European Conference and Exhibition on Geographical Information Systems (EGIS), pages 300-307.

**N. Reddy, B. Prassad, P. Reddy and A. Gupta (1994)** A Methodology for Integration of Heterogeneous Databases, IEEE Transactions on Knowledge and Data Engineering Vol. 6, Num. 6, pages 920-933.

**D. Richard (1993)** Documents de références FEIV 1.1, rapport techniques IGN.

**P. Rigaux (1994)** La représentation multiple dans les systèmes d'informations géographiques, Revue internationale de géomatique, Hermès, Vol. 4, Num. 2 , pages 137-164.

**P. Rigaux (1995)** Interfaces visuelles et multi-représentation dans les bases de données spatiales thèse CNAM Paris.

**P. Rigaux et M. Scholl (1995)** Multi-Scale Partitions : Application to Spatial and Statistical Databases, SSD, Egenhofer Herring (Eds.), Springer-Verlag, Portland, pages 170-183.

**P. Rouet (1991)** Les données dans les systèmes d'information géographique, Hermès, Paris.

**A. Ruas et J-P. Lagrange (1995)** Generalisation : State of the art and issues, GIS and GENERALISATION, GISDATA, Müller, Lagrange, Weibel (Eds.), Taylor & Francis, pages 73-90.

**J. Rumbaugh, M. Blahaet, W. Premerlani, F. Eddy et W. Lorensen (1991)** Object-Oriented modeling and design, General Electric Research and Development center, Prentice Hall, Englewood Cliffs, New Jersey 07632.

**F. Salgé (1995)** Intégration de bases de données géographiques hétérogènes : Le cas des données topographiques, Base de donnée et systèmes d'information pour l'environnement, INRIA, Ministère de l'environnement, E. Simon (Ed.), Versailles.

**E. Salmeron et M. Milgram (1986)**, Utilisation de la relaxation pour la mise en correspondance des segments d'une carte et d'une image aérienne, Semaine Internationale de l'Image Electronique, Deuxième Colloque Image, Nice, avril 1986, pages 32-38.

**F. Saltor, M.G. Castellanos et M. Garcia-Solaco (1992)** Overcoming Schematic Discrepancies in Interoperable Databases, IFIP DS-5 Conference Semantics of Interoperable Databases Systems, Australie, pages 184-198.

**S Scherrer, A. Geppert, K. Dittrich (1993)** Schema Evolution in NO 2, rapport technique Institut fur Informatik der Universitat Zurich.

**M.H. Scholl and M. Tresch. (1993)** Schema Transformation without database reorganisation, ACM SIGMOD Record, vol. 22, Num. 1.

**M.H. Scholl et H-J. Schek (1990)** A relational object model, 3rd Int. Conference On Database Theory (ICDT'90), Paris.

**M. Scholl, A. Voisard, J-P. Peloux, L. Raynal et P. Rigaux (1996)** SGBD Géographiques Spécificités, International THOMSON Publishing.

**G. Schorter, L. Raynal et F. Vauglin (1994)** $GéO_2$ : Module de Superposition, Journées Cassini, Lyon, pages 251-261.

**S. Servigne (1993)** Base de données géographiques et photos aériennes : de l'appariement à la mise à jour, thèse INSA Lyon.

**S. Servigne (1994)** Base de données géographiques et photos aériennes : mise en correspondance, 5th European Conference and Exhibition on Geographical Information Systems (EGIS), Paris, pages 1390-1399.

**K. Shea et R. McMaster (1991)** Cartographic Generalisation in a digital environment When and how to generalize, Map generalization, Buttenfield and McMaster (Eds.), Longman Scientific & Technical, pages 103-118.

**I. Shepherd (1992)** Geographic Information Systems, chapitre 22 : Information Integration and GIS, Maguire, Goodchild, Rhind (Eds.), Longman Scientific & Technical, pages 337-358.

**A. Sheth et V. Kashyad (1993)** So Far « Schematically » yet So Near « Semantically », Interoperable database systems (DS-5), IFIP, Hsiao, Neuhold, Sacks et -Davis (Eds.), pages 283-312.

**A. Sheth et J. Larson (1990)** Federated database systems for managing distributed, heterogeneous, and autonomous databases, ACM Computer Surveys Vol. 22, Num. 3, pages 183-236.

**B. Shmutter et Y. Doytsher (1992)** Matching a set of digitized cadastral maps, CISM Journal ACGC, Vol. 46, Num. 3, pages 277-284.

**P. Shoval et S. Zohn (1991)** Binary-relationship integration methodology, Data and Knowledge Engineering, Elsevier Pub., Vol. 6, Num. 3, pages 225-250.

**S Sleath et S Perry (1996)** Geospatial Object Model Report, rapport technique Ordnance Survey : National Topographic Database Research, num. NTDR/T005.

**P. Souquiere (1994)** La carte électronique : Définitions et principes, Bulletin du comité Français de Cartographie, pages 47-59.

**C. Souza dos Santos (1994)** Design and Implementation of an Object-Oriented View Mechanism, BDA, Clermont-Ferrand.

**C. Souza dos Santos, S. Abiteboul et C. Delobel (1994)** Virtual Schemas and Bases, International Conference on Extending Data Base (EDBT), Base Technology, Cambridge, pages 81-94.

**S. Spaccapietra et C. Parent (1991)** Conflicts and Correspondence Assertions in Interoperable Databases, ACM SIGMOD RECORD, vol.20, num. 4.

**S. Spaccapietra, C. Parent et Y. Dupont (1992)** Model Independent Assertions for Integration of Heterogeneous Schemas, Very Large DataBases Journal, Vol. 1, Num. 1, pages 81-126.

**S. Spaccapietra et C. Parent et T. Devogele (1996)** Analysis of Discrepancies in Spatial Data Representation, CODAS Cooperative Database Systems for Advanced Systems, Kyoto University, ACM Japan, ACM SIGMOD Japan.

**E. Stephan, A. Vckovski et F. Bucher (1993)** Virtual data set : An approach for the integration of incompatible data, Auto Carto 11, ASPRS/ACSM, Minneapolis, pages 93-102.

**N. Stricher (1993)** Base de données multi-échelles : Association géométrique entre la BD Carto et la BD Topo par mesure de la distance de Hausdorff, rapport de stage Paris 11 réalisé au COGIT.

**D. Sui et J. Giardino (1995)** A multi-scale and multi-zoning scheme study for the city of Houston, GIS/LIS.

**C. Thieme et A. Siebes (1993)** Schema Integration in Object Oriented databases, 5th Conference on Advanced Information Systems Engineering CAiSE, Springer-Verlag, pp.54-70.

**S. Timpf et T. Devogele (1997)** New Tools for Multiple Representations, ICC'97, ICA/ACI, Stockholm, pages 1381-1386.

**S. Timpf and A. Frank (1995)** A Multi-scale DAG for cartographic objects, in Auto Carto 12, Charlotte, pages 157-163.

**S. Timpf, G. Volta, D. Pollock et M. Egenhofer (1992)** A conceptual model of wayfinding using multiple levels of abstraction, Theories and methods of spatio-temporal reasoning in geographic space, Goos et Hartmanis (Eds.), Springer-Verlag, pages 348-367.

**J. Trevisan (1995)** Conception d'une BD multi-échelles, rapport de stage ingénieur ENSG réalisé au COGIT.

**T. Ubeda et M. Egenhofer (1997)** Topological Error Correcting in GIS, Advances in Spatial Databases (SSD'97), Berlin, Scholl et Voisard (Eds.), Springer-Verlag, pages 283-297.

**H. Uitermark (1996)** The Integration of Geographic Databases, Second Joint European Conference (JEC), Rumor, McMillan et Ottens (Eds.), IOS Press, pages 92-95.

**U.S. Government (1994)** Coordinating Geographic Data Acquisition and Access : The National Spatial Data Infrastructure. Office of the press Secretary, The White House, USA, Avril.

**F. Vauglin (1997)** Modèles statistiques des imprécisions géométriques des objets géographiques linéaires, Thèse de l'Université de Marne-la-Vallée, Laboratoire COGIT.

**H. Veregin (1989)** Error modeling for the map overlay operation, Accuracy of Spatial databases, M. Goodchild & S. Goptal (Eds.) Taylor & Francis.

**A. Voisard et H. Schweppe**. **(1994)** A Multilayer Approach to the Open GIS Design Problem, ACM GIS workshop, ACM Press, New York.

**J.M. Ware (1994)** Multi-Scale Data Storage Schemes for Spatial Information Systems, thèse University of Glamorgan or The British Library.

**J. Ware et C. Jones (1992)** A multiresolution topographic surface database, Int. Journal on Geographical Information Systems (IJGIS), vol.6, num. 6, pages 479-496.

**C. Weber (1994)** Comment obtenir des données comparables lorsque les règles de recensement changent, 5th European Conference and Exhibition on Geographical Information Systems (EGIS), Paris, pages 604-611.

**M. Worboys et M. Deen (1991)** Semantic Heterogeneity in Distributed Geographic Databases, SIGMOD Record, Vol. 20, Num. 4, pages 30-34.

**B. Yaser (1996)** A Mechanism for object identification and transfer in a heterogeneous distributed GIS, Advances in GIS Research II, Spatial Data Handling (SDH), Delft, Kraak et Molenaar (Eds.), Taylor & Francis, pages 929-941.

**K. Yétongnon, and M. Andersson, and Y. Dupont and S. Spaccapietra (1993)** Conception d'une Base de Données Fédérée, AFCET.

**M Yuan et J. Albrecht (1995)** Structural Analysis of Geographic Information and GIS Operation from a User's Perspective, COSIT : Spatial Information Theory, Vienne, pages 107-122.

**Zhan (1996)** A set of Shortest path algorithms : That run fastest on real road networks, GIS/LIS'96, Denver, pages 755-764.

# 9. Publications

**T. Devogele, C. Parent et S. Spaccapietra (1998)** On Spatial Database Integration, à paraître dans Int. Journal on Geographical Information Systems(IJGIS), vol. 12, Taylor & Francis.

**T. Devogele (1997)** Les SIG Multi-représentations : Applications et Systèmes, Informatique pour l'Environnement Umweltinformatik'97 , Strasbourg, Geiger, Jaeschke, Rentz, Simon Spengler, Zilliox Zundel (Eds.), Metropolis, page 678-687.

**S. Timpf et T. Devogele (1997)** New Tools for Multiple Representations, ICC'97, Stockholm. (Suède), organisé par ICA/ACI, pages 1381-1386.

**S. Spaccapietra, C. Parent et T. Devogele (1996)** Analysis of Discrepancies in Spatial Data Representation, Cooperative Database Systems for Advanced Systems (CODAS), Kyoto (Japon), organisé par l'université de Kyoto, ACM Japan et ACM SIGMOD Japan.

**C. Parent, S. Spaccapietra et T. Devogele (1996)** Conflicts in Spatial Database integration, Parallel and Distributed Computing Systems (PDCS), Dijon, organisé par CACM.

**T. Devogele, J. Trevisan et L. Raynal (1996)** Building a multi-scale database with scale-transaction relationships, Spatial Data Handling (SDH), Delft (Pays-Bas), Kraak et Molenaar (Eds.), pages 6.19-6.33 ou 337-351.

**T. Devogele, J. Trevisan et L. Raynal (1996)** Processus de constitution d'une base de données multi-échelles, Revue Internationale de géomatique, Vol. 6, Num. 2-3, pages 249-263.

**T. Devogele (1996)** Constitution d'une base de données multi-échelles, Bulletin d'information de l'IGN N°65, pages 44-49.

**T. Devogele, J. Trevisan et L. Raynal (1996)** Designing and producing a Multi-Scale Database from Mono-Scale Databases, Geoinformatics, West Palm Beach (Etats-Unis), organisé par CPGIS et SFWMD, pages 14-25.

**T. Devogele et L. Raynal (1996)** Modelling a Multi-Scale Database with Scale-Transaction Relationships, First International Conference on Geographic Information Systems in Urban Regional and Environmental Planning, Sellis T. (Eds.), Samos (Grèce), pages 83-93.

**T. Devogele et L. Raynal (1995)** Intégration de bases de données à différentes échelles : Les liens de correspondances, Secondes Journées de la Recherche CASSINI, Marseille, Jeansoulin R. (Ed.).