ANGELO AUGUSTO FROZZA

UM MÉTODO PARA DETERMINAR A EQUIVALÊNCIA SEMÂNTICA ENTRE ESQUEMAS GML

FLORIANÓPOLIS – SC
2007

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Angelo Augusto Frozza

UM MÉTODO PARA DETERMINAR A EQUIVALÊNCIA SEMÂNTICA ENTRE ESQUEMAS GML

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Dr. Ronaldo dos Santos Mello

Florianópolis, Abril de 2007.

UM MÉTODO PARA DETERMINAR A EQUIVALÊNCIA SEMÂNTICA ENTRE ESQUEMAS GML

Angelo Augusto Frozza

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, Área de Concentração: Sistemas de Computação, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____
Prof. Rogério Cid Bastos, Dr.
Coordenador do PPGCC

Banca Examinadora:

_____
Prof. Ronaldo dos Santos Mello, Dr.
Orientador

_____
Prof. Clodoveu Augusto Davis Junior, Dr.

_____
Prof. Frank Augusto Siqueira, Dr.

_____
Prof[a]. Lia Caetano Bastos, Dra.

_____
Prof. Renato Fileto, Dr.

“A descobrir fatos novos e isolados,

eu preferia ligar fatos já sabidos.”

Alexander von Humboldt

A meu pai, Annibale (*in memoriam*),

que viu este trabalho iniciar,

mas não teve a oportunidade de vê-lo concluído.

A minha mãe, Liduvina,

por sua força e dedicação.

A meu irmão, Alexandre (*in memoriam*),

pois sinto saudades.

AGRADECIMENTOS

Agradeço a Deus pelas oportunidades que me são apresentadas, pela sabedoria para tomar a decisão de aceitá-las e por me dar força para abraçá-las.

A minha mãe, Liduvina, e minha irmã, Lucilia, pelo carinho, apoio e confiança, vocês são muito importantes para mim.

A Márcia, pelo amor, apoio, companheirismo e torcida, pelas trocas de humor, ausências e atrasos que compartilhamos neste período.

Ao meu orientador, Prof. Dr. Ronaldo dos Santos Mello, pelos ensinamentos, conselhos, oportunidades, atenção e confiança dispensados.

A Dona Maria, mãe da Márcia, e ao Marcos, pelo apoio e por me receberem durante minhas viagens a Florianópolis.

Aos meus colegas de trabalho na Universidade do Planalto Catarinense – UNIPLAC, Wilson, Sabrina, Perin e Viviane, pela amizade e apoio.

Aos meus amigos Daniel, Leandro, João e Hernani, pela ajuda, apoio e por estarmos juntos na *IT Factory*.

Aos amigos do Grupo de Banco de Dados da UFSC, em especial a Rodrigo Gonçalves, Fabiano Carniel, Leonardo Rosa, Alexandre Lazaretti, Rafael Vasel e ao Prof. Renato Fileto, pela amizade, convívio, ambiente, *happy hours* e pelas dicas nas reuniões quinzenais do grupo.

Este trabalho não poderia ser concluído sem a participação de inúmeras outras pessoas, que trocaram idéias, atenderam aos meus *e-mails*, forneceram materiais de pesquisa, enfim, ajudaram de inúmeras formas. Alguns foram lembrados nas referências bibliográficas, aos outros peço desculpas por não nomear um a um, pois temo esquecer alguém e o espaço me limita. A todos que de alguma forma contribuíram, deixo meus mais sinceros agradecimentos e a vontade de continuarmos em contato.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ASCII	- *American Standard Code for Information Interchange*
BD	- Banco de Dados
CEP	- Código de Endereçamento Postal
CTM	- Cadastro Técnico Municipal
DAML+OIL	- DARPA *Agent Markup Language+Ontology Inference Layer*
DOM	- *Document Object Model*
FGDC	- *Federal Geographic Data Committee*
GML	- *Geography Markup Language*
GML_I	- Esquema GML importado
GML_M	- Esquema GML principal
IBGE	- Instituto Brasileiro de Geografia e Estatística
INPE	- Instituto Nacional de Pesquisas Espaciais
IPTU	- Imposto Predial e Territorial Urbano
MADS	- *Modeling of Application Data With Spatio-temporal features*
MAV	- *Metrics for Atomic Values*
MCV	- *Metrics for Complex Values*
MGE	- *Modular GIS Environment*
MID	- *MapInfo Interchange File* (arquivo de dados tabulares)
MIF	- *MapInfo Interchange File* (arquivo de dados gráficos)
MUB	- Mapa Urbano Básico
MUB-BH	- Mapa Urbano Básico de Belo Horizonte
OGC	- *Open Geospatial Consortium*
OGOC	- *Ontology-based Geo-Object Catalog*
OMT-G	- *Object Modeling Technique for Geographic Applications*
OWL	- *Web Ontology Language*
OWL-DL	- OWL *Description Logics*
PRODABEL	- Empresa de Informática e Informação do Município de Belo Horizonte
PUC-Rio	- Pontifícia Universidade Católica do Rio de Janeiro
RDF	- *Resource Description Framework*
SDTS	- *Spatial Data Transfer Standard*
SGBD	- Sistemas Gerenciadores de Banco de Dados
SIG	- Sistema de Informações Geográficas
SHP	- *Shapefile*
SPR	- *Spring*
SU	- Sistemas Urbanos
UFRGS	- Universidade Federal do Rio Grande do Sul
UML	- *Unified Modeling Language*
XMI	- XML *Metadata Interchange*
XML	- *eXtensible Markup Language*
XML *Schema*	- *eXtensible Markup Language Schema*
W3C	- *World Wide Web Consortium*

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

RESUMO

Com o crescimento do uso de Sistemas de Informações Geográficas (SIGs) e o alto custo para a produção de dados georreferenciados, o compartilhamento e a troca de dados entre organizações é uma alternativa interessante. Porém, devido à forma como cada aplicação armazena seus dados, a interoperabilidade entre SIGs apresenta incompatibilidades nos níveis sintático e semântico. O objetivo deste trabalho é tratar uma parte específica deste problema: a determinação de equivalências semânticas entre esquemas GML (*Geography Markup Language*). Para tanto, segue a tendência de uso da GML, que é uma linguagem para o transporte e armazenamento de dados geográficos baseada em XML (*eXtensible Markup Language*), e da OWL (*Web Ontology Language*), que é a especificação mais recente para a definição de ontologias recomendada pelo W3C (*World Wide Web Consortium*). Como solução, apresenta um método para determinar as equivalências semânticas entre esquemas GML apoiado por uma ontologia desenvolvida em OWL. O trabalho considera o domínio do cadastro urbano como estudo de caso, uma vez que este domínio é pouco explorado em trabalhos relacionados e apresenta grande potencial de aplicação. O método realiza três grandes atividades: a) criação de uma representação canônica para a ontologia e os esquemas GML envolvidos; b) aplicação de um algoritmo para determinar as equivalências semânticas entre os elementos dos esquemas GML; c) mapeamento das equivalências encontradas entre os elementos dos esquemas GML. O formato de representação canônica escolhido é uma estrutura em árvore, a qual se mostra adequada para representar e manipular dados XML. O algoritmo de determinação de equivalências semânticas divide o problema em pequenas partes, para as quais aplica um conjunto de sete métricas de similaridade, complementadas por um conjunto de pesos e limiares utilizados para definir a importância semântica de cada componente analisado e delimitar o ponto a partir do qual os resultados podem ser considerados corretos. O mapeamento das equivalências é feito automaticamente ou com a participação de um especialista, sendo mantido em um banco de dados relacional. Uma ferramenta para validação do método foi implementada e os experimentos realizados nela geraram um considerável número de equivalências corretas como resultado.

Palavras-chave: Sistema de Informações Geográficas; interoperabilidade semântica; métricas de similaridade; banco de dados geográficos; GML; OWL.

ABSTRACT

With the increase in the use of Geographic Information Systems (GIS) and the high costs of georeferenced data production, data sharing and exchange among organizations is becoming an interesting alternative. Nevertheless, due to the way that each application stores its data, interoperability among GISs faces incompatibilities on both the syntactic and the semantic levels. The objective of this work is to deal with a specific part of this problem: the establishment of semantic equivalences among geographic data schemas. We work on two recommendations: GML, which is a language for geographic data transport and storage based on the same principles of the XML (*eXtensible Markup Language*), and OWL (*Web Ontology Language*), which is the most recent specification for ontology definition recommended by the W3C (*World Wide Web Consortium*). Our solution is a method to determine semantic equivalences among GML schemas supported by a developed ontology in OWL. We focus on the domain of urban cadastre as a case study, became this domain is little explored on related work and has high application potential in the area of urban planning. The method performs three important activities: a) the creation of a canonical representation for the considered ontology and GML schemas; b) the application of an algorithm to determine semantic equivalences among the elements of the GML schemas; c) the mapping of the equivalences founded among the elements of the GML schemas. The adopted format for the canonical representation is a tree structure, which has shown to be suitable to represent and manipulate XML data. The proposed algorithm separates the problem into small parts, to which it applies a set of seven similarity metrics, complemented by a set of weights and thresholds that set the semantic relevance of each analyzed component and delimit the point from which the results can be considered correct. The definitive mappings, defined either automatically or with the assistance of an expert, are kept in a relational database. A tool that supports the method was implemented, and some experiments were performed on it, all of which generating a considerable number of correct equivalences.

Keywords: Geographic Information Systems; semantic interoperability; similarity metrics; geographic database; GML; OWL.

1 INTRODUÇÃO

1.1 Apresentação

Sistemas de Informações Geográficas (SIGs) são sistemas que realizam o tratamento computacional de dados geográficos. Diferente dos sistemas de informações convencionais, os SIGs têm como principal característica a capacidade de armazenar tanto os atributos descritivos como as geometrias dos diversos tipos de dados geográficos (CÂMARA, 2005).

O crescimento da Sociedade da Informação impulsionou o uso de Sistemas de Informações Geográficas nas organizações. Porém, a produção de dados georreferenciados[1] tem um custo muito alto, gerando a necessidade das organizações compartilharem e trocarem seus dados geográficos, o que é facilitado pelas redes de computadores e pela Internet.

Devido à representação complexa da informação geográfica e a falta de padrões estabelecidos nessa área, a troca de informações entre SIGs distintos apresenta diversas incompatibilidades. Tais incompatibilidades se manifestam em dois níveis: sintático e semântico.

O nível sintático se refere ao formato (instâncias, esquema e modelo de dados) de armazenamento e documentação dos dados em cada sistema. Este nível baseia-se na conversão sintática direta de formatos de exportação e importação, sendo o *shapefile* (SHP) o formato mais comum (ESRI, 2005).

O nível semântico, por sua vez, diz respeito ao significado dos dados geográficos em cada SIG. A simples capacidade de transferir dados de um sistema para outro não garante que os dados tenham significado para o novo usuário. A interoperabilidade em SIG requer um nível de modelagem semântica para explicar a correspondência dos conceitos entre diferentes sistemas (CÂMARA *et al*., 1999).

---

[1] Neste trabalho optou-se por grafar a palavra *georreferência* e derivadas com o dígrafo *rr*, por encontrar o vocábulo grafado com dígrafo ou sem na mesma proporção em trabalhos relacionados.

Assim, um dos desafios da comunidade científica é desenvolver soluções que permitam a interoperabilidade entre SIGs, tais como propostas de padrões para intercâmbio de dados e desenvolvimento de SIGs baseados em ontologias (FONSECA, EGENHOFER e BORGES, 2000).

O desenvolvimento de soluções visando à interoperabilidade entre SIGs aponta como tendência o uso de padrões abertos. Nesta linha, o consórcio *Open*GIS (OGC, 2007) tem papel fundamental, por ser uma organização cuja meta é desenvolver especificações de interfaces de dados espaciais para uso global. Dentre os principais trabalhos do *Open*GIS destaca-se a especificação da *Geography Markup Language* (GML) (OGC, 2004).

O objetivo da GML é oferecer um conjunto de regras com as quais um usuário possa definir sua própria linguagem para descrever seus dados. A partir da versão 3.0 (OGC, 2004), a GML consolida-se como um padrão para o transporte e armazenamento de informação geográfica. Esta versão traz muito mais recursos que outras versões, elaborados a partir das necessidades atuais dos usuários e das deficiências encontradas nas versões anteriores, principalmente no que se refere à forma de representar objetos geográficos (SILVA, 2003). Como exemplo de evolução, até a versão 2.1.2 da GML, o modelo de dados era baseado em feições simples (*simple features*), utilizadas para representar fenômenos do mundo real com propriedades geométricas simples, definidas por coordenadas bidimensionais. A versão 3.0 incorporou o conceito de objetos geográficos, sendo uma feição geográfica uma especialização de um objeto geográfico. Essa característica permitiu ampliar as possibilidades de representação geográfica da GML.

Apesar dos avanços encontrados na versão 3.0 da GML, a representação semântica ainda é limitada. Como a GML é definida através de esquemas XML (XML *Schemas*), ela herda as mesmas dificuldades de representação semântica da XML (*eXtensible Markup Language*), ou seja, a XML não possui recursos que permitam reconhecer a semântica de um domínio em particular (SILVA, 2003).

Segundo Lake (2002), para resolver esse problema pode-se incorporar alguma descrição ontológica, como DAML+OIL (DARPA *Agent Markup Language+Ontology Inference Layer*), RDF (*Resource Description Framework*) e OWL (*Web Ontology*

Language), associada a um esquema GML. Com o uso de ontologias procura-se promover a interoperabilidade semântica entre SIGs com diferentes arquiteturas e formatos de dados (PINHO e GOLTZ, 2003).

Ontologias vêm sendo amplamente utilizadas como estratégia de representação do conhecimento sobre um determinado domínio de interesse, fornecendo um esquema semântico que se mostra eficaz, já que permite especificar de maneira explícita e formal os termos do domínio e os relacionamentos entre os mesmos (AZEVEDO *et al*., 2006).

1.2 Definição do Problema

O problema abordado nesta dissertação insere-se no contexto de um problema ainda maior, que é a garantia da interoperabilidade semântica entre esquemas GML.

A determinação de equivalências semânticas entre esquemas GML é uma operação que, dados dois esquemas de entrada, produz um mapeamento entre os elementos desses esquemas que se referem às mesmas classes de entidades do mundo real. Este tipo de operação deve levar em conta as imprecisões existentes na identificação de equivalências semânticas, o que, muitas vezes, não permite definir uma correspondência exata entre dois elementos. Nestes casos, a definição das similaridades de dois elementos é uma alternativa viável, considerando as diversas abordagens baseadas em métricas de similaridade disponíveis na literatura e seus resultados satisfatórios.

O desenvolvimento de ferramentas para a determinação de equivalências semânticas entre esquemas GML é uma das principais etapas para se alcançar a interoperabilidade. Tais ferramentas permitem que os elementos presentes em esquemas distintos possam ser mapeados entre si.

1.3 Justificativa

A necessidade de troca de dados geográficos entre organizações, tanto das produtoras desses dados como daquelas que apenas utilizam dados geográficos, é uma realidade no momento atual. SIGs distintos possuem diferentes formas de representar a realidade geográfica através de formatos de dados próprios. Isso faz com que cada

usuário siga o modelo de dados relacionado ao sistema por ele adotado. Esta diversidade acarreta problemas na troca de dados entre organizações utilizando SIGs distintos, que incluem distorção de dados, comprometimento de qualidade da informação, perda de definições de atributos e georreferenciamento (LIMA, 2002).

Segundo Casanova *et al.* (2005), há um crescente volume de fontes de dados geográficos independentes, o que abre diversas oportunidades para intercâmbio de dados em tempo hábil, reduzindo custos e agilizando processos de decisão. Porém, para se utilizar estas fontes de dados disponíveis, as aplicações devem ser capazes de acessar, interpretar e processar dados provenientes de diversas fontes. De acordo com Hohl (1998), em um ambiente de sistemas heterogêneos, a conversão de dados representa um custo entre 60% e 80% do custo total na implantação de SIGs.

O intercâmbio de dados corresponde à capacidade de compartilhar e trocar informações e processos entre diferentes usuários da informação. O desafio encontra-se na semântica do funcionamento de cada SIG e na maneira como os dados estão organizados, ou seja, no modelo conceitual dos SIGs (CASANOVA *et al.*, 2005).

Como os SIGs não seguem modelos comuns, o compartilhamento de dados geográficos deve envolver processos para garantir que a informação não seja perdida ou corrompida na transferência. Essa não é uma tarefa simples, por causa da complexidade inerente à informação geográfica (CASANOVA *et al.*, 2005).

A abordagem mais básica para compartilhamento de dados geográficos é a conversão sintática direta, que procura traduzir os arquivos de informação geográfica entre diferentes formatos, como SHP (*shapefile*, *ArcView*), MID/MIF (*MapInfo*), SPR (*Spring*) etc. (CASANOVA *et al.*, 2005).

Uma limitação dessa abordagem está nas diferenças de entendimento entre comunidades de usuários distintas. Visões diferentes da realidade geográfica sempre existem, uma vez que as pessoas têm culturas diferentes e a própria natureza é complexa e conduz a percepções distintas (CASANOVA *et al.*, 2005).

Outros esforços para padronização de formatos e uniformização sintática e semântica de objetos geográficos apontam para o uso de padrões abertos. Neste sentido, a XML (*eXtensible Markup Language*) (XML, 2007) aparece como um padrão para a troca de dados. O *Open Geospatial Consortium* (OGC, 2007) fornece um conjunto de

padrões para se trabalhar com dados geográficos, sendo um dos principais a *Geography Markup Language* (GML) (OGC, 2004). Porém, segundo Davis Junior *et al.* (2005), o aspecto semântico na GML não é considerado de forma efetiva à promover a interoperabilidade. Por exemplo, dois usuários de domínios diferentes podem representar uma determinada entidade em GML como <*rio*> e <*curso_de_agua*>, respectivamente. Para a troca de dados entre esses dois usuários, é necessário que também compartilhem os esquemas GML, assim a aplicação que processa a troca pode saber que <*rio*> ou <*curso_de_agua*> são especializações da classe GML <*_Feature*> e pode tratá-las adequadamente. Porém, não há como saber se <*rio*> tem o mesmo significado de <*curso_de_agua*> e vice-versa.

Assim, as pesquisas atuais procuram criar mecanismos que permitam combinar essas diferentes visões que correspondem ao conhecimento geográfico, visando à interoperabilidade pela equivalência semântica dos conceitos existentes em sistemas distintos (CASANOVA *et al.*, 2005). A principal corrente neste sentido propõe o uso de ontologias para representar o conhecimento e a concepção de SIGs baseados em ontologias (FONSECA, EGENHOFER e BORGES, 2000).

Trabalhos relacionados (RODRÍGUEZ, EGENHOFER e RUGG, 1999; MOROCHO, PÉREZ-VIDAL e SALTOR, 2003; BRAUNER, CASANOVA e LUCENA, 2004; HESS e IOCHPE, 2004; STOIMENOV, STANIMIROVIC e DJORDJEVIC-KAJAN, 2004; SOTNYKOVA, CULLOT e VANGENOT, 2005) abordam a interoperabilidade semântica entre SIGs, apresentando diversas abordagens. Uma característica comum a todos é a definição de um ambiente fortemente acoplado, com ênfase na transformação de consultas. Geralmente, as soluções apresentadas abrangem qualquer domínio de aplicação, mesmo quando a validação é feita em um único domínio.

Diferente desses, este trabalho enfatiza a integração de dados geográficos em um domínio de aplicação específico, como sugere Silva (2003). O método proposto considera que as fontes de dados não estão interligadas, mas podem, regularmente, realizar trocas de dados geográficos para atualização de uma base de dados comum. Ainda, procura utilizar os padrões mais recentes, ou seja, a GML 3.1.1 e a linguagem OWL, disponíveis a partir de 2003 (FROZZA e MELLO, 2006a).

1.4 Objetivo Geral

O objetivo geral deste trabalho consiste em apresentar um método para a determinação de equivalências semânticas entre esquemas GML no domínio do cadastro urbano, o qual é pouco explorado em trabalhos relacionados. Este método prevê a utilização de uma ontologia de domínio que serve de base de conhecimento, contendo a descrição dos conceitos utilizados no domínio em questão.

1.5 Objetivos Específicos

Os objetivos específicos são:

- Especificar um processo de determinação de equivalências semânticas entre esquemas GML;
- Detalhar os tipos de equivalência tratados;
- Detalhar as métricas de similaridade utilizadas;
- Desenvolver um estudo de caso para demonstrar e validar o uso do método, descrevendo a ontologia e os esquemas GML utilizados.

1.6 Metodologia e Estrutura do Trabalho

Este trabalho é composto por seis capítulos. Este primeiro capítulo apresenta a motivação para o trabalho, destacando a definição do problema, a justificativa e os objetivos geral e específicos.

O segundo capítulo é dedicado à revisão da literatura. Primeiramente, o problema da interoperabilidade semântica entre SIGs é apresentado de forma mais detalhada. Posteriormente, são apresentados trabalhos relacionados com o tema, abordando soluções já propostas para o problema e identificando as técnicas utilizadas para desenvolver cada solução. Ao final do capítulo são destacadas as bases que permitiram a definição da presente proposta e uma comparação com as abordagens apresentadas em trabalhos relacionados.

O terceiro capítulo é dedicado à delimitação do escopo do trabalho, ou seja, à descrição do domínio do cadastro urbano. Nesse sentido, destaca também a análise das estruturas dos dados que servem de entrada para o processamento do método proposto

no capítulo 4, as quais correspondem a esquemas GML e à ontologia. A análise dessas estruturas foi desenvolvida levando em consideração a origem dos dados e a construção de cada esquema.

O quarto capítulo descreve o método proposto para a determinação de equivalências semânticas entre esquemas GML no domínio do cadastro urbano. Para tanto, apresenta o método de uma forma geral para, em seguida, detalhar os tipos de equivalência abordados e as métricas de determinação de equivalência utilizadas.

O quinto capítulo apresenta um estudo de caso, desenvolvido para demonstrar e validar a proposta. Inicialmente, é feita a especificação dos dados de entrada utilizados, ou seja, dos esquemas GML e da ontologia. Em seguida, a aplicação do método é descrita passo-a-passo, através de um exemplo simples apoiado por uma ferramenta desenvolvida para esse propósito. Ao final, é descrita a validação do método, através da apresentação das medidas de *recall* e *precision*, obtidas dos experimentos realizados.

O sexto é último capítulo apresenta as considerações finais do trabalho e propostas de trabalhos futuros.

2 TEMAS RELACIONADOS

Este capítulo apresenta uma revisão da literatura sobre os tópicos que deram origem à proposta de trabalho. Inicialmente, é apresentado o problema da interoperabilidade entre Sistemas de Informações Geográficas (SIGs) heterogêneos para, em seguida, focar os tipos de conflitos semânticos que podem ocorrer. Na seqüência, são apresentados alguns trabalhos relacionados ao tema, incluindo tanto trabalhos que descrevem soluções completas para a interoperabilidade quanto trabalhos que tratam apenas uma parte do problema. Por fim, são registradas as características gerais que norteiam a proposta de uma solução para o problema da interoperabilidade, relacionando essas características com as definidas para este trabalho.

2.1 O Problema da Interoperabilidade

SIGs são ferramentas computacionais para o processamento de informações sobre a superfície terrestre, auxiliando o ser humano na monitoração, administração e planejamento do espaço geográfico em que vive. Tais sistemas estão inseridos em uma área científica denominada Geoprocessamento (THOMÉ, 1998).

Apesar de sua importância, uma das grandes limitações para um maior uso de SIGs é que a aquisição e a estruturação de dados georreferenciados (geo-dados) são atividades caras e que demandam tempo e pessoal habilitado (FORNARI e IOCHPE, 2002). Este processo envolve algumas atividades de alto custo como, por exemplo, a captura de imagens do terreno a partir de aerofotogrametria ou o uso de satélites.

Um modo de reduzir os custos de aquisição de dados geográficos e prevenir a redundância de esforços é incentivar a troca de informações entre instituições (FORNARI e IOCHPE, 2002). Assim, instituições que já investiram na aquisição de dados geográficos podem obter alguma receita com a comercialização destes dados a um custo menor do que os provedores tradicionais ou disponibilizá-los sem custos para outras instituições. Câmara (2006) defende que, quando os dados públicos são disponíveis de forma aberta, todos ganham. As empresas podem oferecer mais serviços

com menor custo. As demais instituições públicas podem construir bases de dados melhores e mais abrangentes e o cidadão tem acesso a dados que lhe dizem respeito.

Uma vez que se tem disponível um volume crescente de fontes independentes de dados geográficos, isto se torna mais um fator motivador para o intercâmbio de dados. A expansão das redes de computadores também é um fator que leva ao crescimento no uso de geoprocessamento como ferramenta para tomada de decisão em diversas áreas. Neste último caso, a possibilidade de acesso a dados geográficos através da Internet destaca-se como uma solução que minimiza o problema do custo de aquisição de geo-dados, uma vez que propõe o compartilhamento de dados entre diferentes organizações públicas ou privadas.

Neste contexto, os problemas referentes à troca de dados geográficos estão diretamente relacionados à natureza particular dos dados e à modelagem dos mesmos em um SIG. Esses tipos de problema podem ser divididos em dois casos distintos: intercâmbio e interoperabilidade.

O intercâmbio de dados é a prática mais comum e corresponde ao aspecto sintático do problema. Basicamente, refere-se aos processos de importação e exportação, envolvendo a simples conversão de modelos e formatos de dados.

O armazenamento dos dados geográficos em um SIG é organizado em estruturas próprias que descrevem as características dos dados como, por exemplo, as coordenadas dos pontos que formam um polígono que representa geometricamente uma entidade geográfica. As entidades geográficas possuem uma representação geométrica (ou, simplesmente, geometria) e atributos associados. A geometria tem por base as primitivas ponto, linha e polígono, as quais podem ser derivadas para formar estruturas mais complexas (LIMA, 2002).

A interoperabilidade, por sua vez, tem um sentido mais amplo. Interoperabilidade é a capacidade de compartilhar informações e processos entre ambientes computacionais heterogêneos, autônomos e distribuídos (YUAN, 1997). Para alcançar a interoperabilidade devem ser consideradas incompatibilidades de representação, de estrutura e de semântica (THOMÉ, 1998). Na área do geoprocessamento, entretanto, obter a interoperabilidade não é um processo trivial, dada a variedade e complexidade da representação de dados geográficos.

Apesar de o aspecto sintático facilitar a transformação de dados de diferentes sistemas, a semântica, ou seja, a representação conceitual da informação geográfica que cada sistema possui, impõe algumas limitações para esse processo. Essas limitações estão associadas, entre outros motivos, à distância existente entre comunidades de diferentes culturas e história, que acabam por conceituar e interpretar distintamente a mesma realidade geográfica (THOMÉ, 1998).

De acordo com Gahegan (1997), como conseqüência de o significado dos dados espaciais não ser o mesmo em modelos de SIGs distintos, a transformação destes dados de um sistema para outro pode acarretar inconsistências lógicas caso seja levada em conta apenas a geometria. Além da diferença semântica intrínseca ao modelo do sistema, existem outras criadas pelo usuário do sistema em função da sua interpretação de um fenômeno geográfico para fins de modelagem. Como exemplo, pode-se citar o caso de diferentes usuários que podem modelar "*solos*" em um mesmo SIG, mas de maneiras diferentes. Essas diferenças são provenientes do atendimento de propósitos diferentes (ou seja, domínios diferentes) ou de diferentes interpretações ou percepções da realidade dentro do mesmo domínio e, quando da migração dos dados, essas barreiras semânticas devem ser levadas em conta.

Para Fonseca e Egenhofer (1999), a pesquisa sobre interoperabilidade é motivada pela crescente heterogeneidade em sistemas de computação. A pesquisa sobre integração de bancos de dados vem desde a metade dos anos 80 e a interoperabilidade está se tornando uma ciência da integração. Heterogeneidade em SIG não é uma exceção, mas a complexidade e riqueza dos dados geográficos e a dificuldade de sua representação em sistemas de computação criam problemas específicos para a interoperabilidade em SIG.

Lima (2002) divide o problema da interoperabilidade em dois níveis:

- *Nível Sintático*: corresponde ao esquema de codificação, os arquivos e o formato de exportação próprio que cada sistema usa para descrever as entidades (geometria e atributos). Resume-se à forma de escrever o dado;
- *Nível Semântico*: se refere à representação conceitual da informação geográfica em cada sistema, ou seja, o significado do dado geográfico ou sua intenção.

Segundo Hess e Iochpe (2003), o nível semântico da integração abrange a questão da unificação da terminologia utilizada para representar os fenômenos geográficos e os relacionamentos entre eles. Hess (2004) acrescenta, ainda, que conflitos semânticos também aparecem na composição (características) dos fenômenos e no nível de entendimento do conceito representado pelo fenômeno. Nesse sentido, afirma que é necessário criar uma estrutura de organização do conhecimento, tal como um vocabulário controlado, uma taxonomia, um *thesaurus* ou uma ontologia.

Por fim, a interoperabilidade de dados espaciais, segundo Câmara *et al*. (2000), apresenta os seguintes desafios:

- Falta de modelos conceituais comuns, que acarretam problemas na troca de dados entre SIGs distintos;
- Em ambientes de sistemas heterogêneos, a conversão de dados representa um custo entre 60% e 80% do custo total na implantação;
- No caso brasileiro, isto é agravado pela falta de padrões nacionalmente estabelecidos e pela não disponibilidade de ferramentas de baixo custo.

2.2 Tipos de Conflitos Semânticos

A principal exigência de sistemas de integração de dados é que as diferenças entre a sintaxe e a semântica de fontes de dados heterogêneas não precisam estar visíveis para o usuário. Sistemas de Informações Geográficas governamentais prevêem um domínio de aplicação cuja necessidade de integração de dados é obrigatória, especialmente se os dados são mantidos por diversas agências autônomas e devem ser acessados uniformemente (CRUZ e RAJENDRAN, 2003).

Sheth e Kashyap (1992) afirmam que em qualquer abordagem para a interoperabilidade entre bancos de dados distintos, a questão fundamental é identificar objetos que são semanticamente relacionados e, então, resolver as diferenças sintáticas entre eles. Eles tratam a similaridade semântica entre dois objetos através do conceito de *proximidade semântica*. Para tanto, apresentam uma *taxonomia semântica*, que enfatiza as similaridades entre objetos, e a relacionam com uma *taxonomia estrutural*, que enfatiza as diferenças sintáticas (estruturais ou de representação) entre os objetos.

A taxonomia semântica compreende (KASHYAP e SHETH, 1996):

- *Equivalência semântica*: é a mais forte medida de proximidade semântica que dois objetos podem ter. Diz-se que são semanticamente equivalentes quando representam o mesmo conceito ou entidade do mundo real, ou seja, pode-se definir um mapeamento total 1:1 entre dois objetos, em qualquer contexto;
- *Relacionamento semântico*: é um tipo de similaridade semântica mais frágil do que a equivalência semântica. Diz-se que dois objetos são semanticamente relacionados quando existe um mapeamento parcial *n*:1, uma generalização ou uma abstração de agregação entre os domínios de dois objetos;
- *Relevância semântica*: Diz-se que dois objetos são relevantes semanticamente se eles podem ser relacionados um com o outro usando alguma abstração no mesmo contexto. Assim, a relevância semântica é dependente de contexto, isto é, dois objetos são relevantes semanticamente em um mesmo contexto, mas não são em contextos diferentes;
- *Semelhança semântica*: é a mais fraca medida de proximidade semântica e pode ser útil em certos casos como, por exemplo, quando os domínios de dois objetos não podem ser relacionados por qualquer abstração em qualquer contexto. Aqui, a natureza exata da proximidade semântica entre dois objetos é muito difícil de especificar. Neste caso, o usuário pode ser apresentado a extensões de ambos os objetos, definidas como funções dos objetos em relação a um contexto;
- *Incompatibilidade semântica*: expressa a desigualdade semântica, ou seja, a falta de qualquer similaridade semântica não implica automaticamente que dois objetos sejam semanticamente incompatíveis. Estabelecer a incompatibilidade semântica requer afirmar que não há nenhum contexto ou abstração no qual os domínios dos dois objetos possam ser relacionados.

A taxonomia estrutural, por sua vez, compreende as seguintes classes de incompatibilidade (SHETH e KASHYAP, 1992):

- *Problemas de incompatibilidade de domínio*: correspondem às incompatibilidades existentes entre dois objetos quando eles têm diferentes

definições de domínios de atributos semanticamente similares, ou seja, são as incompatibilidades associadas às definições dos atributos;

- *Problemas de incompatibilidade na definição de entidade*: correspondem às incompatibilidades existentes entre dois objetos quando os descritores de entidade usados pelos objetos são parcialmente compatíveis, mesmo quando o mesmo tipo de entidade está sendo modelado;
- *Problemas de incompatibilidade de valores dos dados*: abrangem as incompatibilidades que aparecem devido as diferenças nos valores dos dados de uma mesma entidade apresentada em diferentes bancos de dados, ou seja, os conflitos dependem do estado do banco de dados;
- *Problemas de incompatibilidade em nível de abstração*: estas incompatibilidades aparecem quando duas entidades similares semanticamente são representadas em diferentes níveis de abstração, ou seja, devido aos diferentes níveis de generalidade em que duas entidades são representadas no banco de dados. Também podem aparecer devido à agregação usada nos níveis de entidade e de atributo;
- *Problemas de discrepância esquemática*: podem ocorrer dentro do mesmo modelo de dados e aparecem quando as informações em um banco de dados (BD) correspondem a metadados em outro BD e também dependem do estado do BD.

Considerando que este trabalho enfatiza os conflitos que podem ocorrer entre dois esquemas GML (*Geography Markup Language*) distintos e relacionados e, que em esquemas não são tratadas as instâncias dos dados, então, a atenção recai especificamente sobre as duas primeiras classes apresentadas acima: incompatibilidades de domínio e de definição de entidades.

Tanto a taxonomia semântica quanto a taxonomia estrutural apresentadas acima não são exclusivas de dados geográficos, podendo ser aplicadas a qualquer tipo de dado.

Os tipos de conflito de domínio possíveis, de acordo com Sheth e Kashyap (1992), são:

- *Conflitos de nome*: dois atributos que são semanticamente parecidos podem ter diferentes nomes, ou seja, são sinônimos. O mapeamento entre sinônimos muitas vezes pode ser estabelecido em relação a qualquer contexto;
- *Conflitos de representação de dados*: dois atributos similares semanticamente podem ter diferentes tipos de dados ou representações. Por exemplo, um atributo pode ser definido como um inteiro de 9 dígitos e outro pode ser definido como uma *string*. A conversão entre diferentes tipos de dados pode ser estabelecida em relação a qualquer contexto;
- *Conflitos de escala dos dados*: dois atributos que são similares semanticamente podem ser representados usando diferentes unidades e medidas. Por exemplo, um atributo que represente o valor de um lote pode ser definido em reais e outro em dólares. O mapeamento pode ser expresso em qualquer contexto em termos de uma função ou uma tabela de pesquisa;
- *Conflitos de precisão dos dados*: dois atributos similares semanticamente podem ser representados usando diferentes precisões. Este caso é diferente do anterior, pois permite um mapeamento exato do domínio com maior precisão para o domínio com menor precisão em qualquer contexto, mas não o contrário. Por exemplo, um atributo representando a nota de um aluno de forma numérica (maior precisão) e outro atributo usando letras (menor precisão - A, B, C, D, E);
- *Conflitos de valor padrão*: dependem da definição do domínio dos atributos. O valor padrão de um atributo é aquele valor que o atributo recebe na ausência de maiores informações do mundo real. Neste caso, dois atributos podem ter valores padrões diferentes em bancos de dados distintos. Por exemplo, o valor da idade de um adulto pode ser definido como 18 anos em um banco de dados e 21 anos em outro. Pode não ser possível especificar um mapeamento entre os valores padrões de dois atributos em todos os contextos, entretanto, muitas vezes é possível definir este mapeamento em relação ao mesmo contexto;
- *Conflitos de restrições de integridade de atributos*: dois atributos similares semanticamente podem ser limitados por restrições que não precisam ser

necessariamente consistentes umas com as outras. Por exemplo, o atributo idade em um banco de dados pode ser limitado a "<= *de 18 anos*" e em outro ser limitado a "> *de 21 anos*". Dependendo da natureza das restrições de integridade envolvidas, pode ser possível generalizar as restrições e ter o mapeamento da restrição mais específica para a mais geral.

A Figura 1 apresenta um resumo dos conflitos de domínio citados e os tipos de proximidade semântica associados a cada conflito.

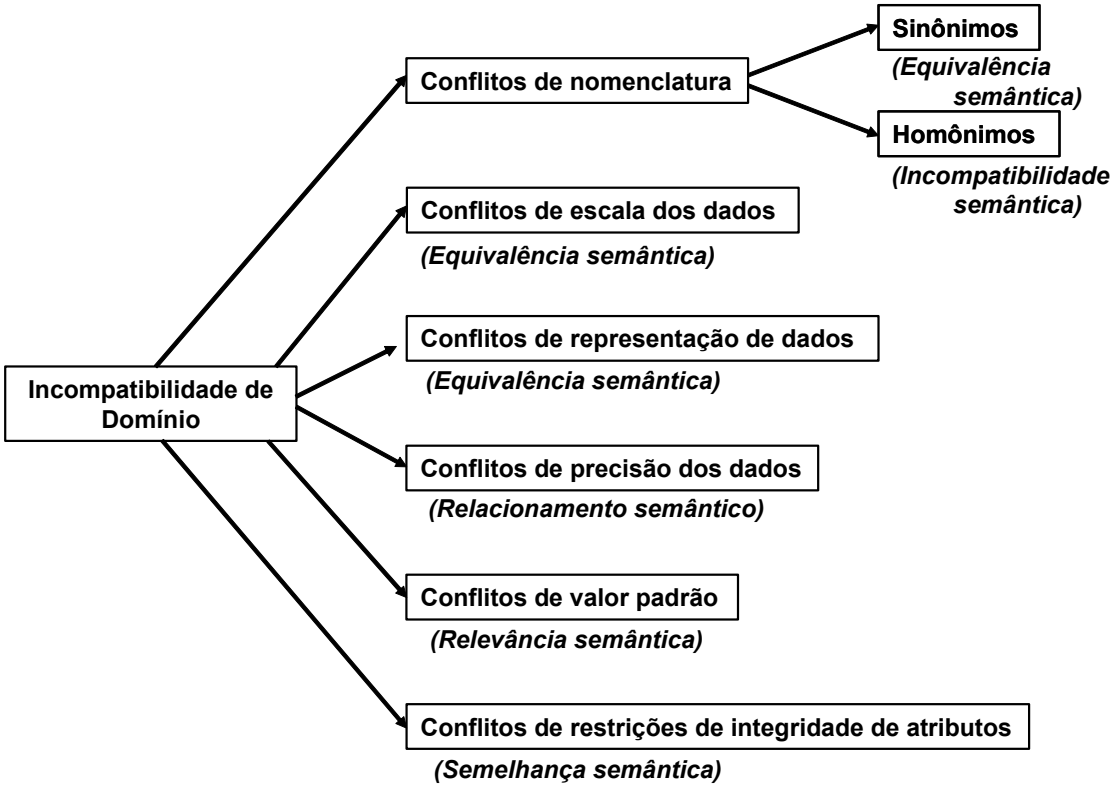


FIGURA 1: Incompatibilidades de domínio e respectivos tipos de proximidade semântica (FONTE: Adaptado de KASHYAP e SHETH, 1996)

Os tipos de conflito de definição de entidade possíveis, segundo Sheth e Kashyap (1992), são:

- *Conflitos de identificador de banco de dados*: as descrições de entidades em dois bancos de dados são incompatíveis porque elas usam registros identificadores que são semanticamente diferentes. Por exemplo, em um modelo relacional, em que duas tabelas representando a mesma entidade têm chaves semanticamente diferentes (em uma é o *nome* e na outra é o

ID). A proximidade semântica dos objetos com este tipo de conflito depende da possibilidade de se definir uma abstração para mapear as chaves de um banco de dados para outro;

- *Conflitos de nome*: entidades podem ser nomeadas diferentemente em bancos de dados distintos. Por exemplo, *Empregados* e *Funcionários* podem ser dois objetos descrevendo o mesmo conjunto de entidades, ou seja, são sinônimos. Tipicamente, o mapeamento entre sinônimos muitas vezes pode ser estabelecido;
- *Conflitos de compatibilidade por união*: descritores de entidades similares semanticamente podem não ser compatíveis por união uns com os outros. Duas entidades são incompatíveis por união quando o conjunto de atributos não é semanticamente relacionado, de tal forma que o mapeamento 1:1 não é possível entre os dois conjuntos de atributos. Por exemplo, uma entidade tem os atributos *Nome* e *Idade*, enquanto a outra tem *Nome* e *Endereço*. Por outro lado, mapeamentos podem ser estabelecidos entre os objetos, tendo por base os atributos comuns;
- *Conflitos de isomorfismo de esquemas*: entidades similares semanticamente podem ter a quantidade de atributos diferentes. Por exemplo, uma entidade tem o atributo *Telefone*, enquanto a outra entidade tem os atributos *TelefoneResidencial* e *TelefoneComercial*. Novamente, as entidades podem ser mapeadas com base nos atributos comuns;
- *Conflitos de item de dado perdido*: ocorre quando duas entidades similares semanticamente têm um atributo perdido. Este tipo de conflito é incluso no mesmo tipo de conflito apresentado anteriormente. Há um caso especial deste tipo de conflito que satisfaz as seguintes condições: a) o atributo perdido é compatível com a entidade; b) existe um mecanismo de inferência para deduzir o valor do atributo.

A Figura 2 resume os tipos de conflito de definição de entidades possíveis e a relação de cada um com a taxonomia semântica.

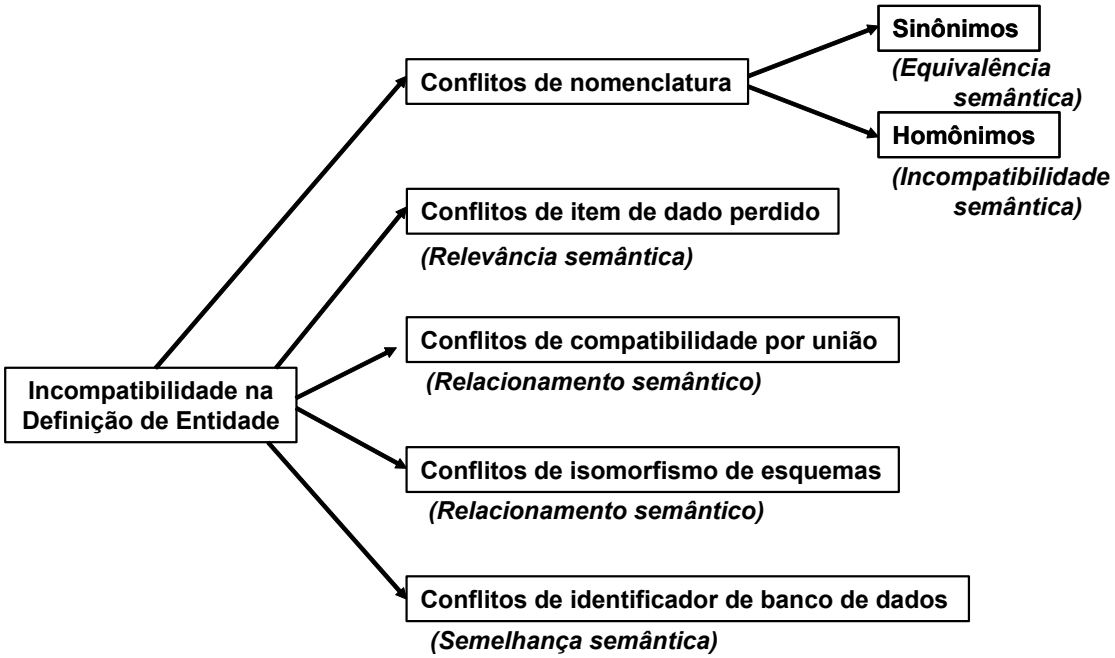


FIGURA 2: Incompatibilidades de definição de entidade e respectivos tipos de proximidade semântica
(FONTE: Adaptado de KASHYAP e SHETH, 1992)

2.3 Propostas para Interoperabilidade de Dados Geográficos

A literatura mostra diversas propostas de integração de dados, desde federações de bancos de dados com esquemas integrados e uso de orientação a objetos, até mediadores e ontologias.

No Brasil, alguns grupos de pesquisa têm se dedicado à pesquisa de soluções visando a interoperabilidade semântica entre SIGs, destacando-se os grupos do Instituto Nacional de Pesquisas Espaciais (INPE), Universidade Federal do Rio Grande do Sul (UFRGS) e Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio).

No INPE, Câmara (2004) defende a criação de um Padrão Nacional de Intercâmbio de Dados Espaciais, destacando em sua proposta:

a) Alternativas existentes para a solução do problema da interoperabilidade:

- Criar um padrão nacional próprio (com base na XML - *eXtensible Markup Language*) ou
- Usar a GML com compartilhamento de esquemas;

b) No caso do uso de XML/GML, é necessária a definição de esquemas para os diferentes tipos de dados públicos como, por exemplo, Censo IBGE – Instituto Brasileiro de Geografia e Estatística, Endereçamento e Cadastro;
c) Disponibilização de ferramentas e dados geográficos de baixo custo.

Thomé (1998) propõe a conversão entre modelos conceituais de diferentes SIGs (*Arc Info*, MGE – *Modular GIS Environment* e *Spring*) para o padrão *Open*GIS através de tabelas de conversão entre as classes de dados que compõem os modelos. Uma das conclusões que apresenta é a fragilidade da versão do *Open*GIS disponível na época, a qual possuía muitos pontos a serem amadurecidos e consolidados e que qualquer interpretação da especificação poderia ser considerada uma "aproximação".

Com a intenção de definir um padrão nacional, Câmara *et al*. (2000) descrevem um formato aberto para ser utilizado na conversão de dados geográficos, baseado em arquivos ASCII (*American Standard Code for Information Interchange*), denominado GeoBR. Este formato supõe que o intercâmbio de dados é baseado em camadas independentes, sendo que cada arquivo GeoBR contém um dado geográfico bem definido (uma *layer*), com todas as informações necessárias para sua decodificação, inclusive com sua descrição (metadados).

Dando continuidade ao trabalho anterior, Lima (2002) transforma a proposta inicial do GeoBR em um modelo genérico para dados geográficos (já como proposta de um padrão brasileiro), com as seguintes características:

- Especifica o GeoBR na forma de um esquema XML, preocupado em atender todo o conjunto de tecnologias de geoinformação;
- Utiliza metadados no cabeçalho do arquivo;
- Inclui, opcionalmente, um dicionário de ontologias para tratar a semântica dos dados geográficos. Este dicionário serve como mecanismo para uma aplicação de equivalência de conceitos.

A comparação entre esta nova proposta do GeoBR e a versão da GML disponível na época (versão 2.1.1) destaca o seguinte (LIMA, 2002):

- O GeoBR utiliza um modelo conceitual único e genérico, não requerendo a definição de esquemas específicos em XML, enquanto a GML requer que cada instituição defina seu esquema de dados;

- O GeoBR tem definições de diferentes tipos de dados (geo-campos e geo-objetos), enquanto a GML 2.1.1 apresenta apenas suporte para geo-objetos simples;
- O GeoBR inclui recursos para intercâmbio em nível semântico.

A GML 3.0 (OGC, 2004), lançada em janeiro de 2003, trouxe importantes atualizações ao padrão, como o suporte a novos tipos de objetos geográficos (3D, superfícies etc.). A partir disso, Silva (2003) analisa esta versão da especificação a fim de descobrir se ela pode ser usada como um padrão aberto que atenda às necessidades brasileiras e, em particular, do INPE. A partir desse estudo, também é reavaliado o formato GeoBR como proposta para formato aberto brasileiro.

O estudo realizado por Silva (2003) reforça a tendência do uso da GML como um padrão aberto para transporte e armazenamento de informação geográfica, apesar das limitações ainda encontradas na versão 3.0:

- Não apresenta uma solução para a questão da interoperabilidade semântica;
- A problemática de integração de dados geográficos codificados em GML e provenientes de fontes heterogêneas ainda persiste, ou seja, pode-se gerar diferentes esquemas GML para representar a mesma informação do mundo real;
- O tratamento da representação temporal apresenta algumas restrições em algumas situações encontradas em aplicações em uso no INPE.

Por outro lado, com base na especificação da versão 3.0 da GML, Silva (2003) aponta que é muito mais difícil e complexo escrever *software* para ler esquemas de aplicação arbitrários, porque os *softwares* devem compreender qualquer conjunto de dados GML. A leitura de um documento GML é trivial, a dificuldade está na interpretação dos elementos XML no contexto geográfico e na interpretação deste contexto geográfico em um sistema local específico. O *software* tem que identificar quais elementos XML representam um objeto geográfico (*feature*), suas propriedades e geometria.

A GML, na sua versão 3.0 (OGC, 2004), é constituída de um conjunto de 33 esquemas-base de aplicação, provendo modularidade aos usuários. Essa versão propõe que, para utilizar os esquemas-base de GML, é necessário desenvolver um esquema de

aplicação específico, associado a um único domínio, ou seja, a GML 3.0 propõe que o usuário somente utilize os esquemas base que são de interesse para sua aplicação. Esta afirmação baseia-se no fato de que um dado geográfico é melhor definido semanticamente em um domínio específico do que pela generalização (SILVA, 2003). Com isso, pode-se dizer, também, que a troca de dados geográficos acontece somente entre domínios que possuem afinidade entre si, como por exemplo, entre aplicações de cadastro urbano ou entre aplicações de cadastro urbano e meio-ambiente (em alguns casos).

Morocho, Pérez-Vidal e Saltor (2003) apresentam um protótipo de ferramenta de integração de esquemas de bancos de dados espaciais, como parte da solução para uma arquitetura federada. Esta ferramenta procura reconhecer as similaridades e diferenças entre entidades a serem integradas. Modelos XMI (XML *Metadata Interchange*) são gerados de representações XML da estrutura dos dados e do banco de dados. Entidades e atributos são extraídos dos modelos XMI por meio de comparações semânticas, para depois avaliar as similaridades e diferenças entre os objetos. Uma ontologia dependente de domínio é criada a partir do padrão FGDC (*Federal Geographic Data Committee*) e de ontologias independentes de domínio (*Cyc* e *WordNet*), sendo representada em OWL (*Web Ontology Language*). Um modelo de proporção (distância entre os nodos da ontologia - *ontology nodes distance*) (RODRÍGUEZ, EGENHOFER e RUGG, 1999; RODRÍGUEZ e EGENHOFER, 2003) é usado para a avaliação das similaridades e diferenças entre termos.

Especificamente sobre a avaliação da similaridade semântica entre definições de classes de feições geográficas, Rodríguez, Egenhofer e Rugg (1999) apresentam uma abordagem que combina duas estratégias diferentes: um processo de casamento de feições e o cálculo da distância semântica. Uma ontologia para o domínio da informação espacial é criada a partir de duas fontes de informação: a *WordNet* e o *Spatial Data Transfer Standard* (SDTS).

Brauner, Casanova e Lucena (2004) propõem o uso de catálogos de geo-objetos baseados em ontologia (*Ontology-based Geo-Object Catalog* - OGOC) aplicados em uma federação de fontes de geo-objetos independentes. Segundo eles,

sistemas remotos devem ser capazes de localizar e acessar fontes de objetos, além de interpretar e processar os objetos. Para tanto, podem ser aplicados três tratamentos distintos a este processo:

- o mapeamento entre pares de esquemas, o qual se torna impraticável se o número de fontes de objetos for razoavelmente grande;
- a adoção de esquemas globais, neste caso, cada esquema conceitual tem de ser mapeado para o esquema global;
- o uso de ontologias para expor o conhecimento implícito, que se apresenta como uma abordagem mais recente.

O trabalho de Brauner, Casanova e Lucena (2004) procura resolver o problema em que duas fontes de objetos geográficos usam identificadores distintos, por exemplo, uma usa o endereço e outra usa um código, sendo que nenhuma das duas armazena os dois identificadores. Assim, sem um mapeamento explícito entre as instâncias de objetos, as duas fontes não podem interoperar. A proposta, então, é que este mapeamento seja feito por meio de um catálogo de objetos. O catálogo age como um mediador para as fontes de dados, provendo serviços para acesso e busca de dados e metadados, armazenando: uma ontologia de referência, similar a um esquema conceitual global; ontologias locais, descrevendo as fontes de objetos; mapeamentos entre as ontologias locais e a ontologia de referência; conjuntos de instâncias de geo-objetos padrão; e mapeamentos de instâncias de geo-objetos de referência para geo-objetos armazenados em cada fonte.

Hess e Iochpe (2004) tratam o problema da equivalência entre esquemas conceituais de bancos de dados geográficos de forma diferente. Em uma abordagem diferente da apresentada por Thomé (1998), foi definida uma ontologia que representa um subconjunto da realidade geográfica e desenvolvido um algoritmo de comparação de similaridades para processar os esquemas contra a ontologia. Nesse trabalho, a GML é usada como um modelo de dados canônico. São definidas regras para mapear outros modelos de dados (GeoFrame, MADS – *Modeling of Application Data With Spatio-temporal features* - e OMT-G – *Object Modeling Technique for Geographic Applications*) para a GML, obtendo-se, assim, um arquivo canônico sem diferenças sintáticas. Uma vez que os modelos estão no formato canônico, passam pelo processo

de tradução semântica, para o qual é utilizada a ontologia. Cada um dos elementos do arquivo GML é comparado aos conceitos na ontologia, a fim de encontrar um compatível. Ao final, é criado um novo esquema no formato canônico, sintática e semanticamente compatível.

Stoimenov, Stanimirovic e Djordjevic-Kajan (2004) vêm desenvolvendo uma solução para o problema da heterogeneidade semântica, denominada *GeoNis*. *GeoNis* é um *framework* para interoperabilidade de aplicações SIG que provê uma infra-estrutura para intercâmbio de dados em um ambiente público local, em que as fontes de dados correspondem a serviços e escritórios locais que produzem geo-dados em algum formato. Cada fonte de informação requer a tradução do fluxo da informação entre a fonte e o sistema *GeoNis*. O fluxo da informação é controlado por meio de mediadores. Assim, o *GeoNis* funciona como um grande gerenciador de acesso a diversas fontes de dados locais, que formalmente especifica a terminologia de cada fonte local usando uma ontologia local. Posteriormente, define a tradução entre cada terminologia local por meio de uma ontologia de alto nível (global). A tradução semântica é desenvolvida para um domínio particular.

Sotnykova, Cullot e Vangenot (2005) focam a integração de esquemas em bancos de dados espaço-temporais. Eles propõem uma abordagem diferente para o problema da interoperabilidade, a qual é dividida em quatro etapas:

a) converter os esquemas das fontes de dados para modelos conceituais MADS;
b) definir mapeamentos entre os esquemas na linguagem de correspondência do MADS;
c) fornecer aos projetistas as soluções estruturais das partes relacionadas nos esquemas;
d) realizar a integração dos esquemas.

A lógica descritiva é usada no final de cada etapa para criar um mecanismo de validação dos resultados, reduzindo a necessidade de intervenção do especialista. O modelo conceitual MADS é usado como modelo padrão de representação dos dados. Para a definição da lógica descritiva, os modelos MADS são traduzidos para OWL-DL (OWL *Description Logics*) com a ajuda de uma ontologia de referência, chamada

MADS-OWL. A validação é feita através do uso de *reasoners* sobre os modelos OWL-DL criados.

2.4 Análise dos Trabalhos Relacionados

Como pode ser observado através dos trabalhos relacionados, a interoperabilidade semântica é um problema atual de pesquisa, motivado pelo crescente número de fontes de dados geográficos disponíveis, e que ainda serão disponibilizados, e pela necessidade de compartilhar os dados entre essas fontes, a fim de reduzir custos e a duplicação de esforços na produção dos dados. A pesquisa sobre a questão semântica dos dados geográficos também é apontada como uma prioridade por centros internacionais (UCGIS, 2002).

Ainda não existe um consenso sobre qual a melhor abordagem para ser utilizada na solução do problema da interoperabilidade, uma vez que essa decisão deve levar em conta diversos fatores como, por exemplo, o domínio da informação, a localização e forma de acesso das fontes de dados, entre outros.

O Quadro 1 apresenta algumas características gerais dos trabalhos relacionados, pelas quais se podem definir algumas linhas de ação.

Como ponto de partida para tratar o problema da interoperabilidade, percebe-se que é necessário ter à disposição os esquemas de dados das diversas fontes de dados. Neste trabalho, parte-se do princípio de que as fontes disponibilizam dados geográficos no formato GML (e seus respectivos esquemas GML) para serem compartilhados. Como esses esquemas são produzidos por pessoas com culturas diferentes, o problema da interoperabilidade permanece.

Assim, para obter a interoperabilidade de fontes de dados em diferentes aplicações que compartilham seus esquemas GML, é necessário definir um modelo de dados que sirva como referência comum, a partir do qual os conceitos dos diferentes sistemas são mapeados. Esse modelo de dados comum pode ser definido na forma de uma base de conhecimento, representada por uma ontologia.

QUADRO 1: Características gerais dos trabalhos relacionados

Referência	Solução proposta	Técnicas utilizadas
Thomé (1998)	Conversão de modelos conceituais de diferentes SIGs (MGE, *Arc Info*, *Spring*) para os conceitos do *Open*GIS. Trata apenas nível sintático.	-Modelos conceituais MGE, *Arc Info*, *Spring;* -Modelo *Open*GIS; -Mapeamento direto.
Rodríguez, Egenhofer e Rugg (1999)	Avaliação de similaridades semânticas entre classes de feições geográficas. Avalia similaridades e diferenças.	-Ontologia (*WordNet* e SDTS); -Métricas de similaridade
Câmara *et al*. (2000)	Definição de um padrão nacional para intercâmbio de dados geográficos – GeoBR.	-Arquivos ASCII; - Formato GeoBR.
Lima (2002)	Expansão do modelo GeoBR e sua especificação como um esquema XML, abrangendo todos os conceitos de um modelo geográfico. Preocupação com o nível semântico (opcional).	-XML *Schema*; -GML 2.1.1; -Ontologias no formato DAML+OIL.
Silva (2003)	Uso da GML 3.0 para intercâmbio de dados geográficos e a extensão do modelo GeoBR para o tratamento de dados espaço-temporais.	-GML 3.0; -Modelo GeoBR.
Morocho, Pérez-Vidal e Saltor (2003)	Uma ferramenta para integração semântica de esquemas de bancos de dados espaciais federados. Avalia similaridades e diferenças. Trata somente a integração de *features* simples. Usa o *Open*GIS como modelo de dados canônico, capaz de representar todos os esquemas com o mínimo de perda de informação dos modelos de dados nativos.	-Arquitetura federada; -Modelos XMI; -Ontologia dependente do domínio: "SIG e BD espaciais"; -OWL; -XML; -GML.
Brauner, Casanova e Lucena (2004)	Propõe um catálogo de geo-objetos (OGOC) para habilitar a interoperabilidade em bancos de dados geográficos federados. Um OGOC cobre um domínio de aplicação específico.	-Ontologia global em OWL; -Ontologias locais.
Hess e Iochpe (2004)	Resolução de heterogeneidades semânticas em esquemas conceituais de BDG dirigida por ontologias. Usa um subconjunto da realidade geográfica (hidrologia). Identifica similaridades e conflitos. Classifica os conflitos semânticos em: igualdade (sinônimos); desigualdade (homônimos); intersecção (igualdade parcial) e pertence (especializações/generalizações).	-Ontologia em RDF - *Resource Description Framework*; -GML 3.0; -Métricas de similaridade.
Stoimenov, Stanimirovic e Djordjevic-Kajan (2004)	Proposta de uma solução para um ambiente heterogêneo, interoperável, distribuído e semântico, baseado em um mediador semântico. A tradução semântica é desenvolvida para um domínio particular (aplicações GIS em serviços públicos locais).	-Ontologias locais; -Ontologia de alto nível (global).
Sotnykova, Cullot e Vangenot (2005)	Uma abordagem diferente para o problema da interoperabilidade: converter os esquemas das fontes de dados para modelos conceituais MADS; definir mapeamentos entre esquemas na linguagem de correspondência do MADS; fornecer as soluções estruturais das partes relacionadas dos esquemas; executar a integração dos esquemas. Entre cada fase, usa a lógica descritiva (OWL-DL) para criar um mecanismo de validação dos resultados, reduzindo a participação do especialista.	-Modelos conceituais MADS; -Lógica Descritiva em OWL-DL.

Outro ponto importante, identificado nos trabalhos relacionados, é a necessidade de se definir um domínio de aplicação. Grande parte dos trabalhos define

"Sistemas de Informações Geográficas" e "bancos de dados geográficos" como domínio de aplicação. Neste trabalho, considera-se que estes domínios são muito abrangentes, uma vez que SIGs podem ser, ainda, subdivididos em diversas áreas de aplicação (meio-ambiente, clima, planejamento urbano etc.), que não necessariamente se relacionam.

Desta forma, optou-se por tratar a interoperabilidade no domínio do cadastro urbano, uma vez que isso pode contribuir para o aumento da troca de dados geográficos no Brasil, considerando que:

- mais organizações vêm tendo interesse em dados geográficos;
- o custo de produção destes dados pode ser compartilhado, evitando-se redundância de trabalho;
- promove-se o desenvolvimento urbano através do planejamento (83% dos municípios brasileiros tem menos de 30.000 habitantes (IBGE, 2000));
- o compartilhamento deve ser maior nos grandes centros (mais pessoas e mais instituições usando os dados).

2.5 Conclusão

Este capítulo foi dedicado ao entendimento do problema da interoperabilidade semântica entre SIGs. Inicialmente, uma abordagem geral sobre o problema da interoperabilidade foi realizada para, em seguida, apresentar um estudo mais aprofundado sobre tipos de conflitos que dificultam a interoperabilidade semântica.

Como o foco deste trabalho são os conflitos que podem acontecer entre dois ou mais esquemas GML e não em instâncias de dados, as incompatibilidades de domínio e de definição de entidades tornam-se o alvo dos estudos. Neste sentido, o capítulo também apresenta os possíveis conflitos relacionados a essas duas classes de incompatibilidades, relacionando-os com os tipos de proximidade semântica que podem ser aplicados a cada caso.

Uma vez obtido o entendimento do problema da interoperabilidade, foi apresentado o estudo de alguns trabalhos relacionados. O objetivo deste estudo foi compreender quais abordagens vêm sendo utilizadas em propostas para solucionar este problema. A partir disso, pôde-se identificar os pontos positivos de cada abordagem e

sua relação com o estado da arte, ou seja, a relação de cada abordagem com as tecnologias disponíveis mais recentes.

A análise dos trabalhos relacionados serviu para delinear as estratégias aplicadas ao método proposto nos próximos capítulos, bem como identificar as tecnologias aplicadas. Nesta etapa destaca-se a intenção de usar as duas principais tecnologias disponíveis atualmente: a especificação 3.1.1 da GML para a criação dos esquemas GML e a OWL como linguagem para definição de ontologias.

A forma como as informações são apresentadas nestas duas linguagens é o assunto do próximo capítulo.

3 DOMÍNIO DE APLICAÇÃO E ESQUEMAS DE DADOS UTILIZADOS

Um ponto fundamental para a solução do problema da interoperabilidade, como salientado anteriormente, é a definição do domínio abordado. Neste trabalho, considera-se o domínio do cadastro urbano, que representa uma das principais aplicações de Sistemas de Informações Geográficas (SIGs).

Inicialmente é feita a descrição do domínio em que o trabalho se enquadra, visando delimitar o escopo abordado. Em seguida, é tratada a criação dos esquemas GML e da ontologia. Ambos representam as informações que o domínio possui e que servem de dados de entrada para processamento pelo método proposto no Capítulo 4.

3.1 Delimitação do Escopo

Com base no fato de que um dado geográfico é melhor definido semanticamente em um domínio específico do que pela generalização (SILVA, 2003), esta seção descreve o domínio do cadastro urbano, o qual foi utilizado como base para o desenvolvimento do método.

O domínio do cadastro urbano foi escolhido para ser abordado neste trabalho porque é pouco explorado em trabalhos relacionados e apresenta um grande número de aplicações práticas possíveis, principalmente aquelas voltadas ao planejamento urbano. A realização de pesquisas sobre planejamento urbano não está limitada somente ao poder público (neste caso, às prefeituras), mas também a outras organizações, como universidades, centros de pesquisa, entre outras. Porém, geralmente são as prefeituras que detêm algum banco de dados (geográfico ou não) de interesse para este tipo de pesquisa. Desta forma, percebe-se facilmente a necessidade de promover a troca desse tipo de dado entre as organizações.

3.1.1 O Cadastro Urbano

Uma prefeitura municipal pode obter muitas vantagens ao investir na utilização do geoprocessamento, apoiado por um SIG como ferramenta para auxiliar o planejamento, a execução e a gestão das políticas públicas municipais. Este tipo de aplicação, também denominada de SIG Urbano (SU), difere de sistemas utilizados em outras áreas e necessita maior precisão e acurácia geográficas, além de preocupar-se com a qualidade da informação alfanumérica (FERRARI, 1997).

SUs aplicados na gestão municipal necessitam trabalhar com um grande número de classes de informações, distribuídas em elementos geográficos, sociais, administrativos e financeiros, entre outros. Eles são utilizados para melhorar a forma de prestação de vários serviços pelos órgãos municipais, em áreas como saúde, educação, trânsito, regulação urbana, saneamento, licenciamento de atividades comerciais e industriais, entre outros.

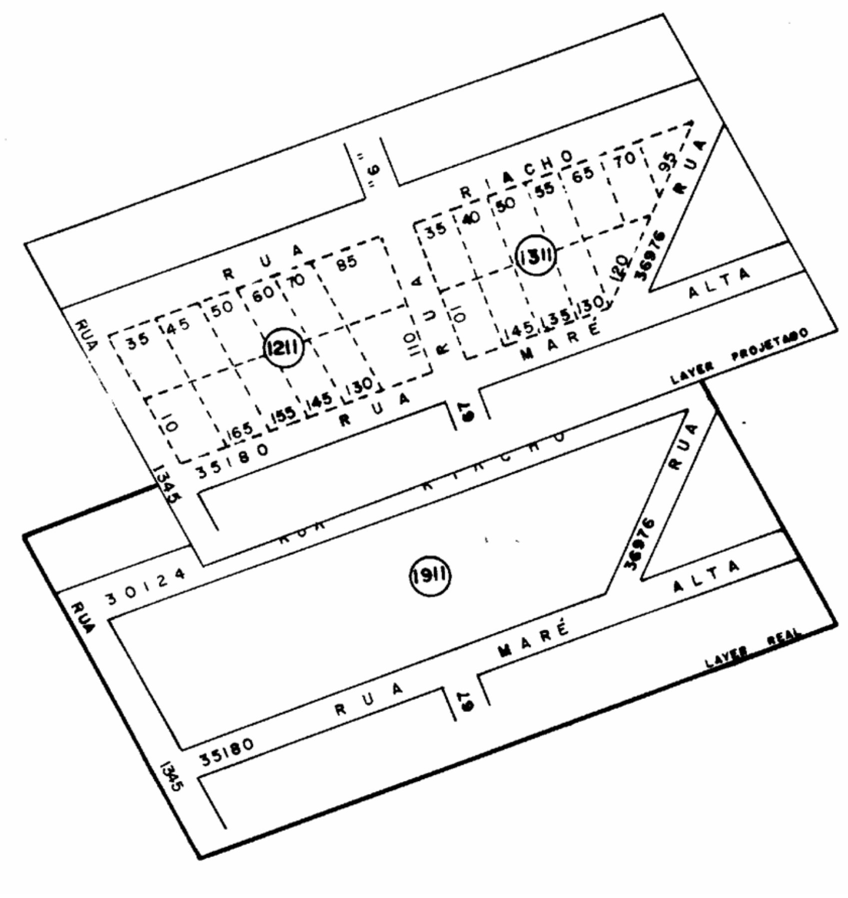


FIGURA 3: Exemplo das camadas *Quadras* e *Lotes*
(FONTE: RIZZO NETO e REIS, 1999)

Como um governo municipal atua em várias áreas, um SIG urbano geralmente é estruturado conforme essa diversidade temática. Além da divisão por áreas (geográficas ou temáticas), pode-se ter, para cada área, as informações dispostas em camadas ou *layers* (Figura 3). Cada camada é composta por um conjunto de classes de objetos geográficos. Exemplos de camadas podem ser: limites oficiais do município, bairros populares, logradouros, quadras, lotes, imóveis etc.

A maioria das tarefas da administração pública municipal é referenciada através de uma estrutura geopolítica dotada de elementos geocodificados, como lotes, arruamentos, quadras, bairros etc. (OLIVEIRA e OLIVEIRA, 2005). A partir da cidade, representada pelos seus elementos essenciais – o espaço geográfico, com suas diversas visões de parcelamento (legal, tributário e real), e todos os objetos que se localizam geograficamente sobre ele, como logradouros, construções, endereços atividades econômicas, elementos de infra-estrutura e demais recursos –, pode-se compreender as dinâmicas de expansão populacional e de uso da terra, permitindo analisar ao mesmo tempo as carências encontradas em cada porção do território (OLIVEIRA e OLIVEIRA, 2005).

O Mapa Urbano Básico (MUB) é o tema central e mais importante em um SIG Urbano, pelo fato de ser necessário para todos os usuários. Ele contempla dados referentes ao "chão da cidade" e reúne classes como: divisões políticas do município, logradouros, quadras, lotes, endereços, eixos de vias, meio fio, pavimentação, hidrografia, parcelamento da terra e diversos outros elementos essenciais. Por ser o principal tema de um SIG Urbano, Oliveira e Oliveira (2005) destacam, inclusive, que há um programa das Nações Unidas voltado à montagem do MUB para municípios. Além disso, este é o tema em que são empregados os maiores investimentos em um SIG, tanto para a formação inicial de um banco de dados geográfico urbano quanto para sua atualização.

Como o MUB serve de suporte para os outros temas, optou-se por concentrar o foco deste trabalho nele. Um exemplo de MUB, específico para a cidade de Belo Horizonte (MUB-BH), é encontrado em Bertini (2003). Neste exemplo, que é aplicado para resolver um problema real, o MUB é representado na forma de diagramas que seguem o modelo conceitual OMT-G (*Object Modeling Technique for Geographic*

Applications). Os elementos que compõem o MUB-BH foram classificados em camadas temáticas (Quadro 2 e Figura 4). Aqui somente são apresentadas as principais camadas e classes. Adicionalmente, o Apêndice II apresenta uma revisão sobre o modelo conceitual OMT-G. Mais detalhes sobre a modelagem do MUB-BH podem ser obtidos em Bertini (2003).

QUADRO 2: Pacotes do MUB-BH

Pacote	Descrição
CTM	Classes do Cadastro Técnico Municipal de Belo Horizonte e seus relacionamentos.
Energia	Classes relativas ao fornecimento de energia elétrica do município e seus relacionamentos.
Hidrografia básica	Classes relativas à hidrografia do município (conjunto de águas correntes ou estáveis) e seus relacionamentos.
Obra pública	Classes oriundas de obras realizadas pelo Poder Público no Município.
Planialtimétrico	Classes relativas à planimetria e altimetria do município.
Telecomunicações	Classe relativa a torres que suportam antenas utilizadas na transmissão de sinais de telefones celulares, TV, rádio etc.
Transporte básico	Classes relativas à infra-estrutura necessária aos meios de transporte (ferroviário e aéreo) do município.

(FONTE: BERTINI, 2003)

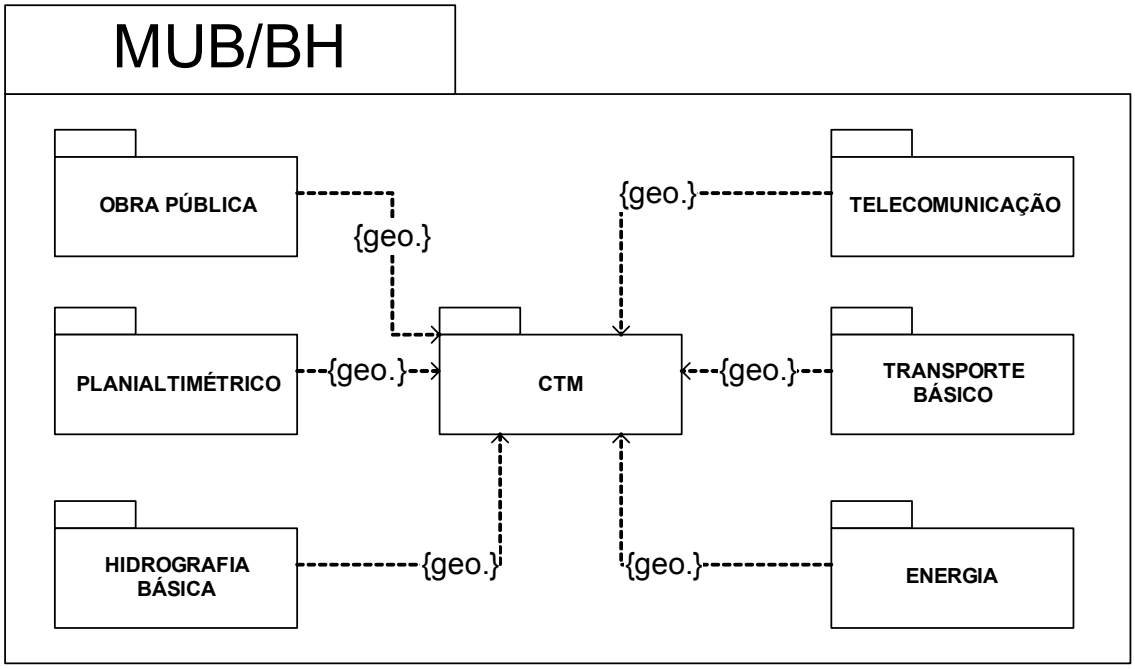


FIGURA 4: Pacotes do MUB-BH
(FONTE: adaptado de BERTINI, 2003)

Conforme ilustrado na Figura 5, o pacote CTM (Cadastro Técnico Municipal) é o principal pacote do MUB-BH, representando as camadas inferiores do Mapa Urbano Básico. Sua composição é descrita no Quadro 3 e na Figura 5. Por necessidade de

restrição de escopo, uma vez que o CTM utiliza um grande número de objetos e conceitos geográficos, este trabalho foca apenas nos pacotes *Lotes* e *Quadras*.

QUADRO 3: Sub-pacotes do pacote CTM

Pacote	Descrição
Aprovações	Classes relativas a projetos e plantas de parcelamento do solo do município.
Edificações	Classes relativas às edificações do município e seus relacionamentos.
Infra-estrutura	Classes relativas à infra-estrutura urbana oferecida aos lotes do município, tais como redes de água, luz, telefone e esgoto, pavimentação, meio-fio etc.
Logradouros	Classes relativas aos logradouros (por exemplo, ruas, praças, avenidas, passarelas, pontes, viadutos, túneis, trincheiras etc.) do município e seus relacionamentos.
Lotes	Classes relativas aos lotes do município (bem como suas divisas, frentes, ocupações etc.) e seus relacionamentos.
Macro divisões	Classes relativas a grandes divisões espaciais da cidade (por exemplo, distritos, regionais, setores, bairros) e seus relacionamentos.
Malha viária	Classes relativas à malha viária (conjunto de vias) do município e seus relacionamentos.
Parâmetros da Lei	Classes relativas aos parâmetros da Lei de Parcelamento, Ocupação e Uso do Solo válidos para o município, tais como áreas de risco, área de tombamento e seu entorno, área de diretrizes especiais, zonas de uso etc.
Quadras	Classes relativas aos quarteirões do município e seus relacionamentos.

(FONTE: BERTINI, 2003)

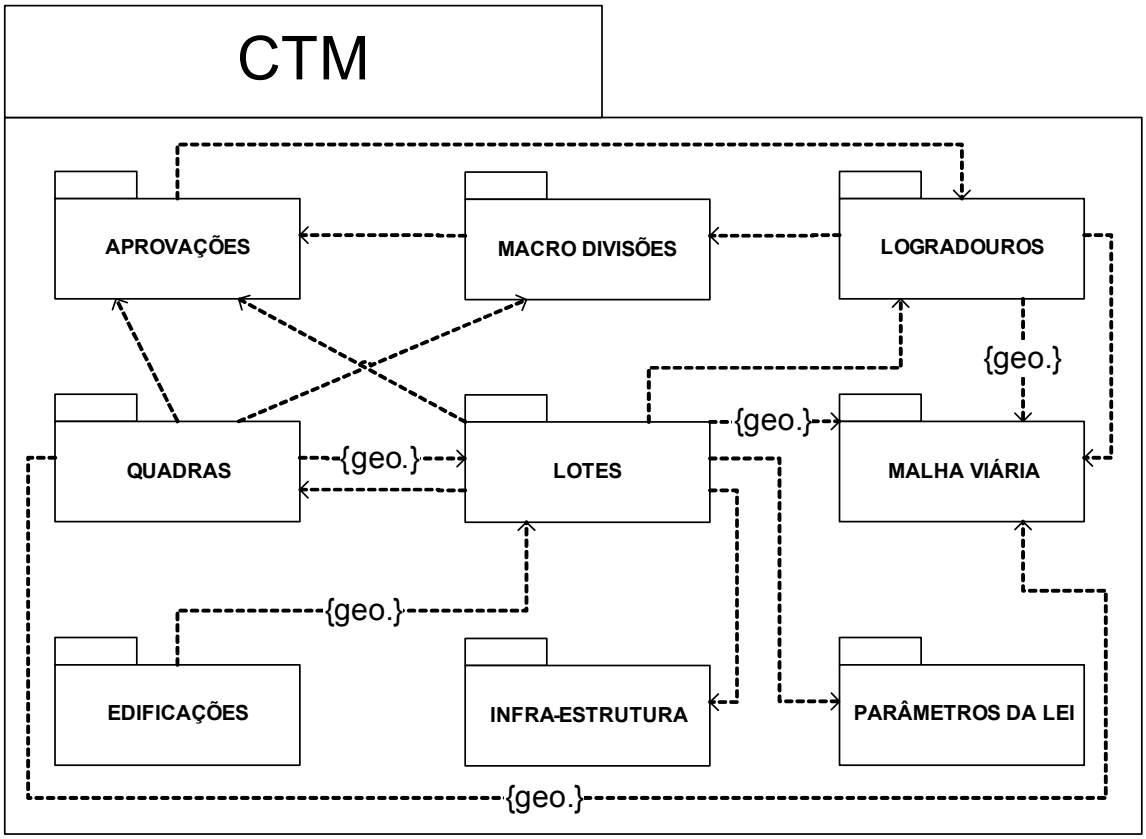


FIGURA 5: Sub-pacotes do pacote CTM
(FONTE: adaptado de BERTINI, 2003)

O Quadro 4 descreve as classes do pacote *Quadras*, incluindo o tipo de representação geográfica de cada classe. Um diagrama OMT-G deste pacote é apresentado na Figura 6.

QUADRO 4: Classes do pacote *Quadras*

Pacote	Descrição	Representação
Quadra_CTM	Polígono fechado regular ou irregular que representa parcela ou divisão do terreno, porção de terra delimitada por logradouros públicos, vias férreas, curso d'água (perene), acidentes geográficos, limite de setor, limite de loteamento (desmembramento de gleba) em quadras com perímetro superior a dez mil metros ou limite de município. É o mesmo que quarteirão. Superfície terrestre constituída de um ou mais imóveis, inserida em um setor cartográfico e devidamente delimitada nas plantas de CTM.	Polígono
Quadra_Projetada	Quadra CTM projetada, não física.	Polígono
Quadra_Real	Quadra CTM fisicamente delimitada.	Polígono

(FONTE: BERTINI, 2003)

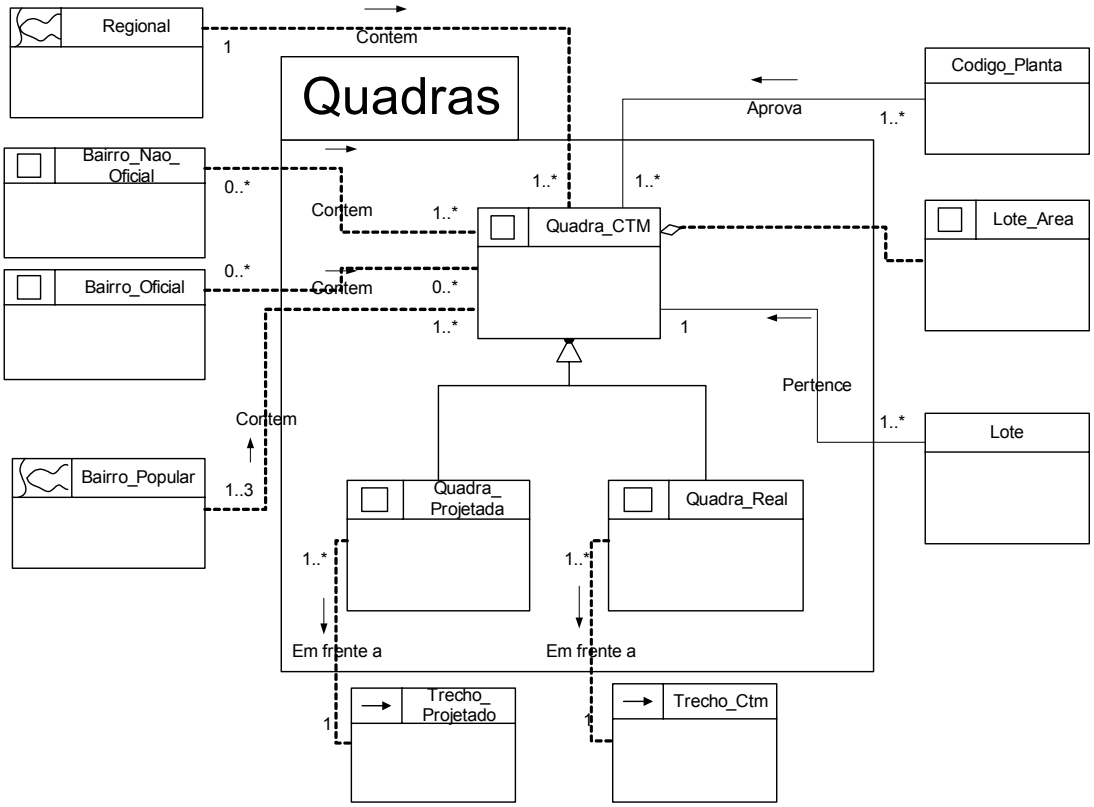


FIGURA 6: Esquema OMT-G para o pacote *Quadras*
(FONTE: BERTINI, 2003)

O Quadro 5 apresenta as classes do pacote *Lote*, com a respectiva representação de cada classe. O esquema OMT-G deste pacote é apresentado no

Apêndice III. Por necessidade de simplificação, neste diagrama OMT-G não estão representados os relacionamentos existentes com outros pacotes.

QUADRO 5: Classes do pacote *Lote*

Pacote	Descrição	Representação
Cemiterio	Identifica a localização dos cemitérios. Regiões ocupadas por grandes equipamentos de interesse municipal ou a elas destinadas.	Polígono
CEP	Código de Endereçamento Postal.	Alfanumérico
Cerca	Elemento com o qual se circunda ou fecha-se um terreno. Pode ser de madeira ou de madeira e arame.	Linha
Compatibilizacao	Indica a equivalência entre os lotes tributável, legal e CTM, cada qual com sua própria chave que o identifica.	Alfanumérico
Divisa_Lote	Qualquer objeto geométrico que caracteriza os vários limites de um lote.	Linha
Endereco	Identificação do imóvel.	Ponto
Lote	Toda porção de terreno edificado ou vago, pertencente a uma determinada quadra. Pode ocupar todo um quarteirão. Pode ser de esquina (com frentes para dois ou mais logradouros que se cruzam), de frente (apresenta frente para um ou mais logradouros públicos), desde que não seja de esquina, de fundo ou lote encravado (sem frente para o logradouro público).	
Lote_Area	Lote identificado como área.	Polígono
Lote_CTM	Porção de terreno definida a partir da codificação de uma situação de direito, identificada de três maneiras: através da planta aprovada ou particular; através do índice cadastral do IPTU; através de visita ao local e pesquisa documental, abrangendo todo o território do município.	Alfanumérico
Lote_CTM_Projetado	Lote ainda não existente fisicamente no CTM, mas para o qual há um projeto de criação.	Alfanumérico
Lote_CTM_Real	Lote CTM fisicamente criado.	Alfanumérico
Lote_Legal	Porção de terreno aprovada e identificada como lote em planta de parcelamento do solo aprovada.	Alfanumérico
Lote_Testada	Lote identificado pela sua frente.	Linha
Lote_Tributavel	Porção de terreno na qual se pode identificar um fato gerador de tributo e resultante de parcelamento oficial, ou de parcelamento informal desde que lançado como lote até 1990.	Alfanumérico
Muro	Muro ou grade.	Linha
Ocupação_Imovel	Ocupação do imóvel levantada no percurso urbano.	Ponto
Testada_Lote	Segmento que compõe o perímetro do alinhamento de um lote voltado para uma via ou confluência de vias; limite divisório entre o lote e a via pública ou a confluência de vias, caracterizando o acesso ao mesmo.	Linha
Testada_Principal	Aquela para a qual o endereço foi oficialmente concedido, identificada nesta ordem: pelo endereço, pela infra-estrutura, pelo grau de hierarquização da classificação viária ou pelo tamanho da testada, conforme apresentada nas plantas do CTM. No caso de haver mais de uma testada é considerada como testada principal, para fins de planta de valores, aquela relativa ao logradouro que confere ao imóvel maior valor.	Linha
Testada_Secundaria	Frente de lote não considerada como principal para fins de planta de valores.	Linha

(FONTE: BERTINI, 2003)

3.2 Criação de Esquemas GML

Segundo Silva (2003), para a construção de aplicações baseadas em XML (*eXtensive Markup Language*) é primordial analisar os requisitos do domínio de aplicação e descrever estes em um modelo de dados. No caso, o modelo de dados em questão é o modelo da GML, que é uma gramática XML definida através de XML *Schema*.

Para o desenvolvimento do método proposto no capítulo seguinte e a validação do mesmo, foi necessário criar um esquema GML para servir de base ao estudo, uma vez que não foram encontrados esquemas GML referentes ao domínio do cadastro urbano disponíveis publicamente.

Na criação do esquema base foi utilizado como referência o trabalho de Bertini (2003), no qual é apresentada a modelagem conceitual do Mapa Urbano Básico de Belo Horizonte (MUB-BH), feita através do modelo conceitual OMT-G (descrito no Apêndice II).

Foi necessário, ainda, fazer a tradução da modelagem conceitual proposta por Bertini (2003) para os respectivos elementos de um esquema GML. Como o modelo conceitual OMT-G segue a formatação geral de um diagrama de classes da UML (*Unified Modeling Language*), a tradução de OMT-G para GML seguiu as regras de mapeamento UML para GML presentes em Hess (2004), com algumas adaptações necessárias em função de se estar utilizando a especificação 3.1.1 da GML (OGC, 2004).

Apesar de esta estratégia estar sendo aplicada ao esquema do MUB-BH, a mesma pode ser aplicada a qualquer esquema modelado em OMT-G ou outro modelo compatível com a UML.

Salienta-se que a GML é constituída por um conjunto de 33 esquemas base de aplicação. Tal especificação provê modularidade aos usuários, permitindo que estes utilizem somente os esquemas necessários à sua aplicação (OGC, 2004).

Ainda que o tema Cadastro Urbano represente um subconjunto do domínio Sistemas de Informações Geográficas, ele apresenta um número relativamente grande de elementos e conceitos. Portanto, foi necessário limitar ainda mais o escopo do

presente trabalho, sendo escolhido neste momento apenas a modelagem referente às classes *Quadras* e *Lotes*, por estes serem os elementos básicos de um Mapa Urbano Básico (MUB).

O método de determinação de equivalências semânticas proposto neste trabalho prevê a utilização de dois esquemas GML:

- um esquema que representa o modelo de dados do sistema principal, ou seja, aquele que recebe os dados dos outros sistemas (GML_M);
- um esquema que representa o modelo de dados que está sendo importado para o sistema principal (GML_I).

O esquema $GML_{M,}$ que já está disponível no sistema principal, também deve passar pela determinação de equivalências semânticas em uma primeira execução do método, para que seus elementos possam ser mapeados contra os conceitos presentes na ontologia.

Os conceitos presentes no modelo OMT-G tratados no presente trabalho podem ser divididos em seis tipos:

- *pacotes* (ou temas);
- *classes*;
- *atributos*;
- *associações*;
- *herança*;
- *aspectos geométricos*.

O mapeamento de outros tipos de elementos presentes no modelo conceitual (campos geográficos, aspectos temporais, aspectos dinâmicos, aspectos de redes e aspectos topológicos ou relacionamentos espaciais) não é mostrado aqui, uma vez que estes outros elementos não são tratados no trabalho por causa da limitação do escopo. Mais detalhes sobre o mapeamento destes elementos podem ser encontrados em Hess (2004) e OGC (2004).

3.2.1 Pacotes

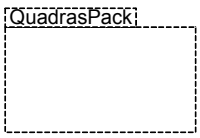
Um pacote (ou tema) existente no modelo conceitual OMT-G representa uma interface que serve para agrupar classes e interfaces relacionadas, organizando e

separando grandes partes de um sistema. Segundo a especificação da GML (OGC, 2004), o mapeamento padrão é gerar um arquivo de esquema GML para cada pacote, porém, em alguns casos é possível colocar mais de um pacote no mesmo arquivo.

Um pacote é mapeado em GML como sendo um elemento global (*element*), cujo atributo *name* é o próprio nome do pacote e o atributo *type* é um tipo de dado definido pelo usuário. O tipo de dado definido pelo usuário recebe o mesmo nome do pacote acrescido do sufixo "*Type*" e deve ser considerado um tipo complexo (*complexType*), de conteúdo complexo (*complexContent*). Este tipo deve estender a classe básica *AbstractFeatureCollectionType* da GML (Figura 7).

Todos os componentes do diagrama que compõem o pacote (classes, outros temas e relacionamentos ou associações) são membros dessa coleção. Assim, complementa-se a definição do pacote com um elemento global que estende a classe *gml:_Feature*. Este elemento é usado para associar os componentes do diagrama aos respectivos pacotes.

A Figura 7 apresenta a codificação de um pacote em GML e o construtor para os membros do pacote.




```
<element name="QuadrasPack" type="urb:QuadrasType"
  substitutionGroup="gml:_FeatureCollection"/>
<complexType name="QuadrasPackType">
  <complexContent>
     <extension
        base="gml:AbstractFeatureCollectionType"/>
  </complexContent>
</complexType>

<element name="_QuadrasFeature" type=
  "gml:AbstractFeatureType" abstract="true"
  substitutionGroup="gml:_Feature"/>

<complexType name="QuadrasPropertyType">
  <complexContent>
     <restriction base="gml:FeaturePropertyType">
        <sequence minOccurs="0">
          <element ref="urb:_QuadrasFeature"/>
        </sequence>
        <attributeGroup ref=
        "gml:AssociationAttributeGroup"/>
     </restriction>
  </complexContent>
</complexType>
```

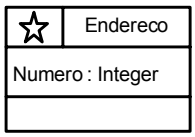
FIGURA 7: Exemplo de mapeamento OMT-G→GML para o pacote *Quadras*

3.2.2 Classes

Cada classe do modelo conceitual OMT-G é convertida para um elemento (*element*) global no esquema GML, com o atributo *name* igual ao próprio nome da classe e o tipo definido pelo usuário. O nome desse tipo é formado pelo nome da classe, adicionando o sufixo "*Type*". Ele deve ser do tipo complexo (*complexType*) e de conteúdo complexo (*complexContent*).

As classes são consideradas especializações de *AbstractFeatureType* e herdam os atributos *fid*, *name*, *description* e *boundedby*. Para associar uma classe ao seu respectivo pacote, na definição da classe o atributo *substitutionGroup* é especificado com o nome do elemento que representa os componentes do pacote ao qual a classe pertence.

A Figura 8 exemplifica a codificação de uma classe em GML.




```
<element name="_Endereco"
  type="gml:AbstractFeatureType" abstract="true"
  substitutionGroup="urb:_LotesFeature"/>
<element name="Endereco" type="urb:EnderecoType"
  substitutionGroup="urb:_Endereco"/>

<complexType name="EnderecoType">
  <complexContent>
     <extension base="gml:AbstractFeatureType">
        <sequence>
          <element ref="gml:Point"/>
          <element name="numero" type="integer"/>
        </sequence>
     </extension>
  </complexContent>
</complexType>
```

FIGURA 8: Exemplo de mapeamento OMT-G→GML para a classe *Endereco*

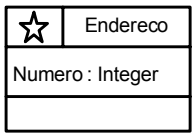
A definição separada do elemento que define a classe do tipo de dado complexo que a representa permite o reuso do tipo de dado quando necessário.

3.2.3 Atributos

Um atributo representa qualquer característica de um objeto. Em GML, os atributos são representados em um tipo de dado complexo através de declarações de elementos XML e não pela declaração de atributos XML.

Assim, cada atributo de uma classe do modelo OMT-G é mapeado em GML como sendo um elemento (*element*) local ao tipo complexo definido para a classe à qual o atributo pertence. O tipo do atributo se mantém o mesmo.

A Figura 9 destaca a codificação de atributos em uma classe GML.




```
<element name="_Endereco"
  type="gml:AbstractFeatureType" abstract="true"
  substitutionGroup="urb:_LotesFeature"/>
<element name="Endereco" type="urb:EnderecoType"
  substitutionGroup="urb:_Endereco"/>

<complexType name="EnderecoType">
  <complexContent>
     <extension base="gml:AbstractFeatureType">
        <sequence>
          <element ref="gml:Point"/>
          <element name="numero" type="integer"/>
        </sequence>
     </extension>
  </complexContent>
</complexType>
```

FIGURA 9: Exemplo de mapeamento OMT-G→GML para atributos

Uma convenção adotada em GML é que os nomes de classes e de tipos de dados iniciam com a primeira letra em maiúsculo, enquanto os nomes de atributos iniciam com minúsculo.

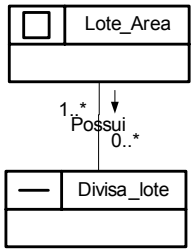
3.2.4 Associação

Cada função (*role*) de uma associação é mapeada para um elemento local (atributo GML) no tipo complexo que define o objeto, com nome igual ao nome da função (*role*) da associação (Figura 10).

O atributo GML que representa uma *role* recebe como tipo uma *PropertyType* que segue o padrão *AssociationType* (OGC, 2004). A indicação da função inversa de uma função (*role*) da associação é feita através de uma anotação *appinfo*.

A Figura 10 apresenta um fragmento da codificação GML para a associação *LoteArea possui DivisaLote* / *DivisaLote pertence a LoteArea*.

Relacionamentos espaciais (disjunto, contém, dentro de, toca, sobrepõe etc.) não são suportados pela especificação atual da GML (3.1.1), porém há previsão de sua disponibilidade para a versão 4 (HESS, 2004).



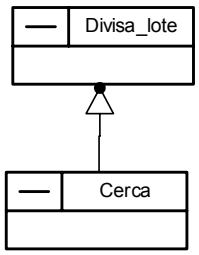

```
<complexType name="LoteAreaType">
   <complexContent>
      <extension base="urb:LoteType">
        <sequence>
           <element ref="gml:Polygon"/>
           ...
           <element name="possui"
              type="urb:DivisaLotePropertyType">
              <annotation>
                <appinfo source="urb:x-
                   gml:reverseProperty">pertence
                </appinfo>
              </annotation>
           </element>
           ...
        </sequence>
      </extension>
   </complexContent>
</complexType>

<complexType name="LoteAreaPropertyType">
   <sequence>
      <element ref="urb:LoteArea" minOccurs="1"
        maxOccurs="unbounded"/>
   </sequence>
   <attributeGroup ref="gml:AssociationAttributeGroup"/>
</complexType>

----------

<complexType name="DivisaLoteType">
   <complexContent>
      <extension base="gml:AbstractFeatureType">
        <sequence>
           ...
           <element name="pertence"
              type="urb:LoteAreaPropertyType">
              <annotation>
                <appinfo source="urb:x-
                   gml:reverseProperty">possui
                </appinfo>
              </annotation>
           </element>
        </sequence>
      </extension>
   </complexContent>
</complexType>

<complexType name="DivisaLotePropertyType">
   <sequence>
      <element ref="urb:DivisaLote" minOccurs="0"
        maxOccurs="unbounded"/>
   </sequence>
   <attributeGroup ref="gml:AssociationAttributeGroup"/>
</complexType>
```

FIGURA 10: Exemplo de mapeamento OMT-G→GML para uma associação

3.2.5 Herança (Generalização/Especialização)

Uma hierarquia de classes do modelo conceitual OMT-G (generalização/especialização) é convertida para uma hierarquia de tipos em GML, em

que cada classe do modelo é codificada como um elemento. O tipo do elemento filho é sempre uma extensão baseada no tipo do elemento pai, ou seja, as classes especializadas são declaradas como extensão da classe genérica. Desta forma, os atributos da classe pai passam automaticamente para as classes filhas. Cabe lembrar que herança múltipla não é permitida na GML, uma vez que cada tipo pode estender apenas um outro tipo. A Figura 11 apresenta um exemplo de codificação para herança de tipos GML.



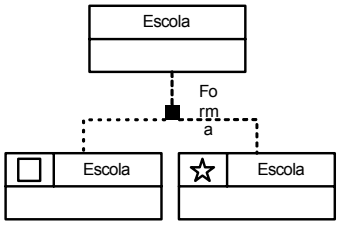

```
<complexType name="DivisaLoteType">
  <complexContent>
     <extension base="gml:AbstractFeatureType">
       <sequence>
         <element ref="gml:LineString"/>
       </sequence>
     </extension>
  </complexContent>
</complexType>

<element name="_Cerca" type="gml:AbstractFeatureType"
  abstract="true"
  substitutionGroup="urb:_LotesFeature"/>
<element name="Cerca" type="urb:CercaType"
  substitutionGroup="urb:_Cerca"/>
<complexType name="CercaType">
  <complexContent>
     <extension base="urb:DivisaLoteType">
       <sequence>
         <element ref="gml:LineString"/>
       </sequence>
     </extension>
  </complexContent>
</complexType>
```

FIGURA 11: Exemplo de mapeamento OMT-G→GML para herança

3.2.6 Aspectos Geométricos (Generalização Conceitual)

No modelo conceitual OMT-G, os aspectos geométricos aparecem na forma de estereótipos (símbolos gráficos) aplicados às classes espaciais do modelo. No mapeamento para GML, este estereótipo é convertido para um atributo que usa uma das propriedades GML definidas para as formas geométricas.

Quando houver a ocorrência de múltiplas geometrias para uma determinada classe, descrevendo diferentes aspectos geométricos da mesma (generalização conceitual, na terminologia OMT-G), essa característica é mapeada como uma escolha (*choice*) dentre as representações possíveis, uma vez que na instanciação somente pode haver uma geometria.




```
<complexType name="EscolaType">
  <complexContent>
     <extension base="gml:AbstractFeatureType">
        <choice>
          <element ref="gml:Point"/>
          <element ref="gml:Polygon"/>
        </choice>
     </extension>
  </complexContent>
</complexType>
```

FIGURA 12: Exemplo de mapeamento OMT-G→GML para aspectos geométricos

No exemplo da Figura 12, um tipo *Escola* é descrito com uma propriedade que indica um ponto que é o centro da Escola e outra que indica a área em volta da Escola.

Uma limitação encontrada no trabalho de Bertini (2003) é que apenas a modelagem das classes e relacionamentos presentes no Mapa Urbano Básico de Belo Horizonte é apresentada e, conseqüentemente, não são indicados os atributos que cada classe deve possuir. Como a informação de quais atributos existem em cada classe é importante para o contexto deste trabalho, estes foram definidos de forma aleatória, obtidos através da leitura de outros trabalhos e documentos relacionados ao MUB de Belo Horizonte (RIZZO NETO e REIS, 1999; GOMES, 2000; OLIVEIRA e OLIVEIRA, 2005).

3.3 A Ontologia de Domínio

O método de determinação de equivalência semântica proposto neste trabalho é dependente de uma ontologia de domínio. Porém, assim como ocorreu com os esquemas GML, não foram identificadas ontologias para o domínio do cadastro urbano disponíveis publicamente.

Como não é objetivo deste trabalho o projeto de uma ontologia de domínio completa para o cadastro urbano, definiu-se apenas a estrutura de uma ontologia e os conceitos principais relacionados com o escopo apresentado na seção 3.1. Assim, tomou-se como base para a criação da ontologia o modelo conceitual de ontologia de lote urbano apresentado por Pinho e Goltz (2003), o qual foi representado em *Web Ontology Language* (OWL) através do editor *Protégé*-OWL (PROTEGE, 2007). A Figura 13 apresenta uma visão simplificada da hierarquia de classes da ontologia. Uma visão completa da ontologia, incluindo seus relacionamentos é apresentada no Apêndice

IV. A OWL é empregada por ser a mais recente especificação de linguagem ontológica recomendada pelo *World Wide Web Consortium* (W3C).

FIGURA 13: Representação simplificada da ontologia

Conforme colocado por Noy e McGuinness (2001), uma ontologia é uma descrição formal explícita de *conceitos* (também chamados *classes*) em um domínio de interesse, incluindo suas *propriedades* (também chamadas *slots*, *roles* ou *funções*) descrevendo as características e atributos dos conceitos, e *restrições* nas propriedades (também chamadas *facetas* ou *restrições de funções*). Uma classe pode ter *subclasses* que representam conceitos mais específicos.

Inicialmente, as classes definidas na ontologia utilizada neste trabalho representam apenas dados discretos (visão de objetos) e dados não-espaciais. A consideração de dados contínuos (visão de campos) é alvo de trabalhos futuros.

A ontologia (ilustrada no Apêndice IV) possui cinco classes principais, três classes complementares e mais dezoito classes especializadas, totalizando 26 classes (Figura 13). Uma descrição das classes é apresentada no Quadro 6.

QUADRO 6: Descrição das classes da ontologia

Classe	Descrição
Conceitualizacao	Superclasse que agrega as formas como um lote é percebido pelos diversos órgãos municipais e pela população. Cada especialização é uma visão complementar do lote e nenhuma tem todos os atributos necessários para acompanhar a complexidade da realidade prática.
Cadastral	(ou real) Corresponde ao lote físico, ou seja, lote enquanto porção do terreno implantado e delimitado no local. Não considera a situação oficial do terreno.
Juridico	Representa o lote que tem documentação registrada em cartório referente à propriedade e características do mesmo.
Legal	(ou oficial) Representa o lote que consta em plantas aprovadas pela prefeitura. Reflete uma visão oficial por ter seu registro em cartório e outras interpretações documentais.
Tributavel	(ou tributário) Lote decorrente do Código Tributário Nacional, que possui comprovação da sua existência e propriedade. É o fato gerador do Imposto Territorial Urbano, componente do IPTU.
LimiteFisico	Superclasse que agrega os tipos de extremidades delimitadoras de um lote.
Cerca	Classe que define um objeto do tipo cerca.
Muro	Classe que define um objeto do tipo muro.
Logradouro	Toda superfície destinada à circulação pública e aos veículos, que permite acesso aos imóveis, às instalações, às áreas de recreação etc.
Lote	Porção do terreno parcelado, com frente para via pública e destinado a receber edificação. Unidade territorial e imobiliária elementar do espaço urbano, onde está alicerçada a economia e a administração municipal.
Edificado	Representação de um lote que possui alguma edificação.
Comercial	Edificação para uso comercial.
Industrial	Edificação para uso industrial.
Misto	Edificação para uso misto.
Residencial	Edificação para uso residencial.
Multi-familiar	Edificação para uso residencial para mais de uma família (condomínios).
Uni-familiar	Edificação para uso de uma única família (casa).
Servicos	Edificação para uso do setor de serviços.
NaoEdificado	Representação de um lote sem edificações.
Quarteirao	Agrupamento de imóveis territoriais (lotes) separados por determinados delimitadores.
RepresentacaoGeografica	Superclasse complementar que agrega as classes das representações geométricas.
Linha	Classe complementar que define um objeto do tipo linha.
Poligono	Classe complementar que define um objeto do tipo polígono.
Ponto	Classe complementar que define um objeto do tipo ponto.
Dicionario	Classe complementar que mantém a lista de sinônimos.
CoordXY	Classe complementar que define um ponto de coordenada (XY)

(FONTE: GOMES, 2000; OLIVEIRA, 2001; PINHO e GOLTZ, 2003)

Na representação da ontologia em OWL foi usada a seguinte convenção para nomes:

- O nome de uma classe inicia com a primeira letra em maiúsculo;
- O nome de uma propriedade inicia com a primeira letra em minúsculo.

Dessa forma, fica fácil para quem estiver interpretando a ontologia diferenciar classes e propriedades. Essa convenção também é útil para uso com o *Protégé* que, diferente de outros editores de ontologia, mantém um único *namespace* para todos os conceitos em uma ontologia e é *case sensitive*, ou seja, diferencia letras maiúsculas de minúsculas.

Um modelo OWL basicamente é composto de classes, propriedades (atributos e relacionamentos) e instâncias. Para representar o conhecimento sobre o domínio, a ontologia utilizada neste trabalho usa apenas a definição de classes e propriedades. A representação de uma classe na ontologia (que corresponde a um tipo *objeto*) é feita através do construtor *<owl:Class>*. Uma classe pode ter subclasses ou superclasses, o que representa uma noção de hierarquia (generalização e especialização). Uma subclasse é definida pelo construtor *<rdf:subClassOf>*. A Figura 14 demonstra a definição em OWL das classes *Quarteirao*, *Logradouro* e *Lote* e a definição da especialização da subclasse *Edificado*.

```
<owl:Class rdf:ID="Quarteirao"/>
<owl:Class rdf:ID="Logradouro"/>

<owl:Class rdf:ID="Edificado">
   <rdfs:subClassOf>
      <owl:Class rdf:ID="Lote"/>
   </rdfs:subClassOf>
</owl:Class>
```

FIGURA 14: Fragmento de código OWL definindo classes e subclasses

Uma *propriedade* OWL é um relacionamento binário que liga dois conceitos OWL, chamados *domínio* (*domain*) e *abrangência* (*range*) da propriedade (SOTNYKOVA, CULLOT e VANGENOT, 2005). Ligações entre duas classes ou entre uma classe e um tipo de dado podem ocorrer, definindo dois tipos de propriedades:

- Propriedades objeto, ligando indivíduos à indivíduos;
- Propriedades tipo de dados, ligando indivíduos a valores de dados (*data values*).

Por padrão, propriedades OWL são multivaloradas, a menos que haja uma restrição explícita. A Figura 15 mostra a definição em OWL de um relacionamento multivalorado (*possuiConceitualizacao*) e de um relacionamento monovalorado (*ehParte*) para o domínio em questão.

```
<owl:ObjectProperty rdf:ID="possuiConceitualizacao">
   <rdfs:range rdf:resource="#Conceitualizacao"/>
   <rdfs:domain rdf:resource="#Lote"/>
</owl:ObjectProperty>

<owl:FunctionalProperty rdf:ID="ehParte">
   <rdfs:range rdf:resource="#Quarteirao"/>
   <rdfs:domain rdf:resource="#Lote"/>
   <rdf:type
        rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
</owl:FunctionalProperty>
```

FIGURA 15: Fragmento de código OWL definindo relacionamentos entre classes

```
<owl:FunctionalProperty rdf:ID="area">
   <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#float"/>
   <rdfs:domain rdf:resource="#Lote"/>
   <rdf:type
     rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
</owl:FunctionalProperty>

<owl:ObjectProperty rdf:ID="centroide">
   <rdfs:domain rdf:resource="#Lote"/>
   <rdfs:range rdf:resource="#Ponto"/>
   <rdf:type
     rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
</owl:ObjectProperty>
```

FIGURA 16: Fragmento de código OWL definindo uma propriedade simples e uma propriedade complexa

Atributos de uma classe podem ser simples ou complexos, monovalorados ou multivalorados, obrigatórios ou opcionais. Um atributo de uma classe é definido em OWL por um dos dois tipos de propriedade descritos acima, dependendo do tipo do atributo. A Figura 16 demonstra a definição em OWL de uma propriedade simples monovalorada (*area*), cuja abrangência (*range*) é um tipo de dado *float*, e uma propriedade complexa monovalorada (*centroide*), cuja abrangência é o conjunto de objetos da classe *Ponto*. A diferença entre a definição de uma propriedade monovalorada e uma propriedade multivalorada é a existência do axioma *FunctionalProperty* no primeiro caso. No caso de propriedades multivaloradas, elas são definidas como *ObjectProperty*, tendo apenas a indicação do *domain* e do *range*.

Um resumo dos elementos estruturais usados em OWL e o correspondente mapeamento para GML é apresentado na Tabela 1.

TABELA 1: Mapeamento entre elementos estruturais da OWL e da GML

OWL	**GML**
Classe/Conceito (*owl:Class*)	Elemento GML
Relacionamento IS-A - generalização/especialização (*owl:SubClassOf*)	Derivação de tipos por extensão
Outros relacionamentos entre classes (*owl:ObjectProperty*, com a definição de *rdfs:range*, *rdfs:domain,* e *owl:inversePropertyOf*)	Propriedades em um elemento GML, com referência a um tipo de dado representando cada função (*role*) do relacionamento
Atributo do tipo Objeto simples (*owl:DatatypeProperty*)	Definição de uma propriedade em um elemento GML (uma propriedade pode ser um atributo ou um relacionamento)
Atributo do tipo Objeto complexo (*owl:ObjectProperty*)	
Atributo identificador (*owl:functionalProperty*)	

De forma semelhante ao que aconteceu com o esquema GML base, o trabalho de Pinho e Goltz (2003) também não apresenta os atributos que as classes da ontologia devem ter. Apenas as classes e seus relacionamentos foram modeladas. Assim sendo, a mesma estratégia usada no esquema GML base foi adotada para a ontologia, ou seja, através da leitura de outros trabalhos relacionados ao Mapa Urbano Básico de Belo Horizonte foram obtidas as definições de atributos apresentadas na ontologia (RIZZO NETO e REIS, 1999; GOMES, 2000; OLIVEIRA e OLIVEIRA, 2005).

3.3.1 A Classe Dicionário

A classe *Dicionário*, presente na ontologia, não representa um conceito particular do domínio do cadastro urbano, mas é utilizada para manter a lista dos sinônimos dos conceitos presentes na ontologia.

Os sinônimos são armazenados como instâncias da classe *Dicionario* (Figura 17). Um banco de dados relacional poderia ter sido utilizado para esta tarefa. Entretanto, optou-se por deixar os sinônimos na ontologia como forma de centralizar em um único local o conhecimento sobre o domínio. De forma semelhante, em vez de criar em cada classe uma propriedade específica para armazenar os sinônimos, preferiu-se deixá-las todas em um mesmo local, facilitando também a manutenção e a pesquisa.

Cada conceito presente na ontologia tem uma instância associada a ele na classe *Dicionario*. O identificador dessa instância tem o mesmo nome da classe,

precedido pelo caractere sublinha ("_"). Uma propriedade do tipo *string*, multivalorada, armazena os sinônimos do conceito, sendo que, para fins de simplificação da implementação, o próprio nome da classe também é tratado como um sinônimo. Além disso, é indicado em que idioma o sinônimo é utilizado, a fim de permitir a escolha da métrica para definição de similaridade de nomes mais adequada para cada idioma.

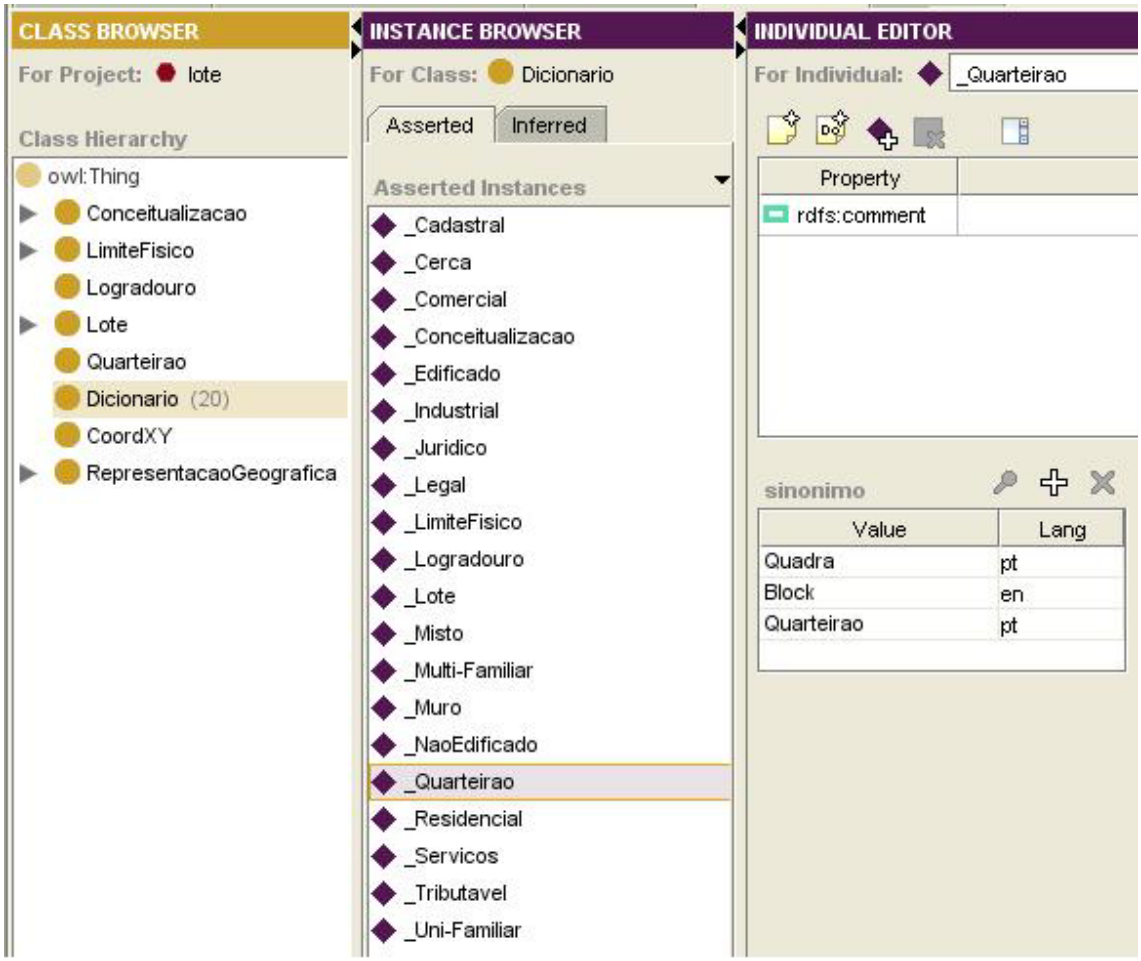


FIGURA 17: Exemplos de instâncias da classe *Dicionario*

3.4 Conclusão

Este capítulo descreve o domínio do problema abordado no trabalho e os esquemas usados como entrada para o método proposto no próximo capítulo. Quanto ao último item, foi abordada a estrutura de esquemas GML e de ontologias OWL.

O potencial de aplicações que podem ser desenvolvidas foi o fator motivador para a escolha do domínio do cadastro urbano como estudo de caso, principalmente aquelas associadas ao planejamento urbano. Por outro lado, esse é um domínio que

possui um grande número de conceitos relacionados, o que levou à necessidade de delimitar o estudo a uma pequena parte desse domínio, ou seja, aos conceitos de *Quadra* e *Lote*. Analisando o domínio do cadastro urbano e classificando seus conceitos por camadas (*layers*), percebe-se que os conceitos de *Quadra* e *Lote* fazem parte da primeira camada, ou seja, são usados como base para as demais. Isto demonstra a importância desses dois conceitos no contexto do domínio e justifica, também, a escolha dos mesmos.

Como não foi possível obter esquemas GML reais e disponíveis publicamente, a solução foi criar os esquemas para uso no trabalho. A estratégia para a criação dos esquemas GML foi, primeiramente, obter os esquemas de bancos de dados geográficos representados em diagramas que seguissem o padrão orientado a objetos de diagramas de classe da UML. Neste sentido, destaca-se o trabalho de Bertini (2003), que usou diagramas OMT-G para representar o MUB da cidade de Belo Horizonte. A partir desse esquema conceitual foi criado o primeiro esquema GML, adaptando-se as regras de tradução UML para GML encontradas em Hess (2004) e OGC (2004). A adaptação das regras foi necessária, a fim de atualizá-las de acordo com a especificação mais recente da GML (3.1.1). Cabe ressaltar que essas regras não são específicas para o esquema do MUB-BH e podem ser usadas com qualquer esquema OMT-G ou que siga o padrão UML. Posteriormente, outros esquemas GML foram criados usando a mesma estratégia, os quais são apresentados no capítulo 5.

De forma semelhante aos esquemas GML, também não foi identificada uma ontologia pronta, disponível publicamente, que abrangesse os conceitos de um Mapa Urbano Básico. Então, a partir de um diagrama conceitual de ontologia para o conceito *Lote*, encontrado em Pinho e Goltz (2003), foi implementada uma ontologia em OWL, usando-se a ferramenta *Protégé*-OWL (PROTEGE, 2007), em sua versão *beta* 2. Foi escolhido o uso da OWL por esta ser a especificação mais recente de linguagem para a descrição de ontologias, recomendada pelo W3C *Consortium*.

Assim, os trabalhos de Bertini (2003) e Pinho e Goltz (2003) foram usados como base para a criação do primeiro esquema GML e da ontologia, respectivamente. Os dois trabalhos acima têm uma característica em comum: foram desenvolvidos usando como estudo de caso a cidade de Belo Horizonte e problemas reais encontrados

na PRODABEL – Empresa de Informática e Informação do Município de Belo Horizonte (http://www.pbh.gov.br). Em função disso, o esquema GML criado a partir do trabalho de Bertini (2003) tem sido tratado como esquema principal (GML_M) no restante deste trabalho.

Outro detalhe em comum dos trabalhos acima citados é que nenhum deles apresenta a definição dos atributos usados, apenas as classes são modeladas. Como a definição de atributos é uma característica importante e necessária para a continuidade deste trabalho, foi preciso consultar outras referências e aleatoriamente definir atributos tanto para o esquema GML quanto para a ontologia. A ausência dos atributos nos trabalhos relacionados ao MUB de Belo Horizonte é justificada, uma vez que grande parte deles está organizada em tabelas convencionais, no âmbito dos sistemas de informações legados (como os sistemas de IPTU, ISS e outros), que apenas recentemente têm sido integrados ao SIG.

4 UM MÉTODO PARA A DETERMINAÇÃO DE EQUIVALÊNCIAS SEMÂNTICAS ENTRE ESQUEMAS GML

Este capítulo apresenta um método para determinação de equivalências semânticas entre esquemas GML (*Geography Markup Language*). A equivalência semântica é obtida calculando-se o grau de similaridade que os elementos dos esquemas GML têm entre si, usando como base o conhecimento do domínio da aplicação armazenado em uma ontologia.

Primeiramente, é apresentada a visão geral do método para, então, serem detalhadas suas três partes principais: pré-processamento, determinação de equivalências e catalogação do mapeamento. O próximo capítulo descreve um estudo de caso que mostra a aplicação do método aqui definido.

4.1 Visão Geral

A visão geral do método é apresentada na Figura 18. O método considera a existência de dois esquemas GML distintos: um esquema representando os dados do SIG principal (GML_M), previamente mapeado para a ontologia, e um esquema representando os dados que estão sendo importados de um segundo SIG (GML_I) (FROZZA e MELLO, 2006b).

O método prevê uma comparação em três passos principais:

- *Passo 1*: criar uma representação canônica para a ontologia de domínio e o esquema GML importado (GML_I);
- *Passo 2*: determinar a equivalência semântica entre o GML_I e a ontologia de domínio;
- *Passo 3*: realizar o mapeamento das equivalências do GML_I com o GML_M.

O método adota uma abordagem de integração de dados binária, ou seja, analisa e determina a integração de esquemas GML aos pares. Esta abordagem foi empregada visando manter o processo o mais simples possível. Para a obtenção das

equivalências semânticas, o método considera a estrutura dos elementos que compõem os esquemas GML e as classes na ontologia, as nomenclaturas utilizadas, os tipos de dados dos atributos e os relacionamentos existentes.

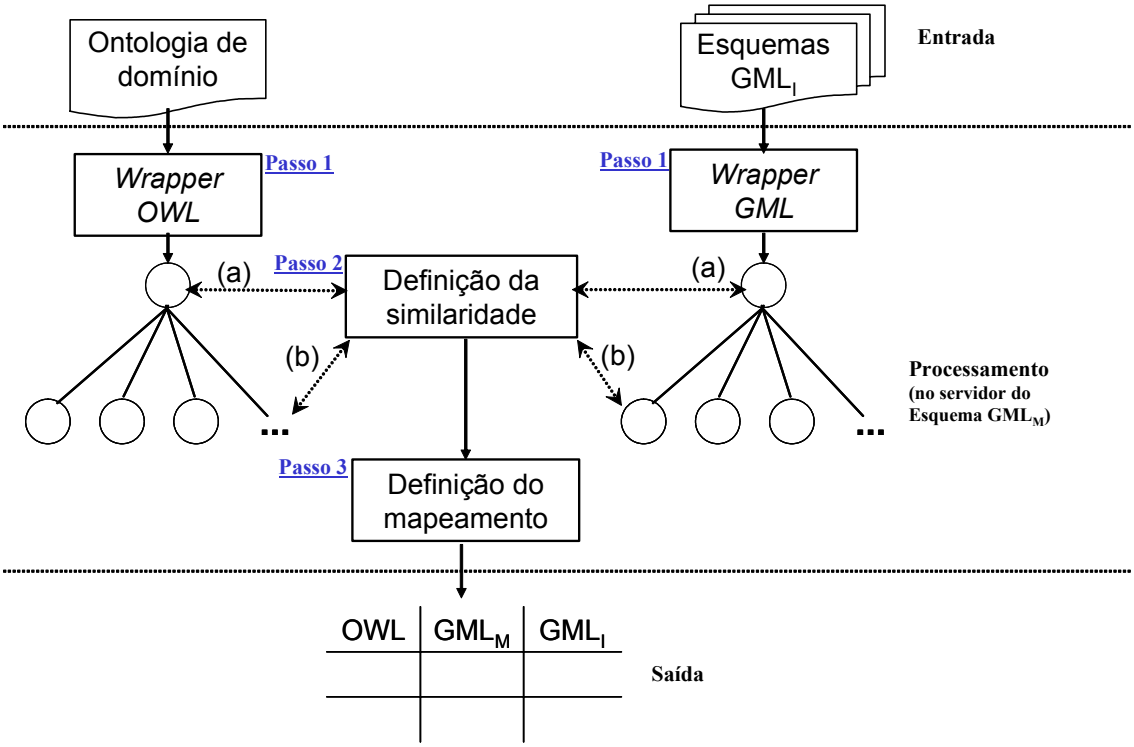


FIGURA 18: Visão geral do método

Os componentes do método podem ser classificados em: componentes de entrada, componentes de saída e módulos internos. Os componentes de entrada compreendem uma ontologia de domínio e os esquemas GML. Os módulos internos correspondem a dois *wrappers* para criar a representação canônica da ontologia e dos esquemas GML, um conjunto de métricas para determinação de equivalência semântica e um algoritmo que define os passos executados pelo método. Os componentes de saída compreendem o módulo de definição do mapeamento e a estrutura do banco de dados. A ontologia representa a base de conhecimento sobre os termos usados no domínio da aplicação. As métricas implementam as regras para identificar se dois dados são similares ou não.

4.1.1 Definições Gerais

Uma ontologia geográfica pode conter tanto conceitos geográficos como não-geográficos (convencionais). Além disso, um conceito possui diversas propriedades

(características), que são os seus atributos e os relacionamentos que este conceito tem com outros conceitos. De forma semelhante, esquemas GML contêm definições de elementos que representam um dado geográfico ou não, os quais também possuem diversas propriedades, ou seja, atributos e relacionamentos.

Algumas definições adaptadas de Hess, Iochpe e Castano (2006) são apresentadas a seguir. Essas definições são úteis para entender como são definidos os objetos (elementos, atributos, relacionamentos etc.) presentes tanto na ontologia quanto nos esquemas GML e, também, para subsidiar a definição da estrutura canônica criada através dos *wrappers* (explicada na seção seguinte).

Como, no contexto desse trabalho, conceitos da ontologia e elementos de um esquema GML possuem estrutura similar (um identificador e um conjunto de atributos e relacionamentos), as definições abaixo apresentadas usam o termo *elemento* para referir-se tanto a conceitos da ontologia como a elementos de esquemas GML. Posteriormente, nas referências às árvores canônicas da ontologia e dos esquemas GML, cada nodo de uma árvore canônica também é referenciado como um *elemento da árvore*.

Definição 1 (Elemento):

Seja O uma ontologia e C o conjunto de conceitos dessa ontologia. Da mesma forma, seja E um esquema GML e C' o conjunto de elementos nesse esquema.

Um elemento $c \in C$ ou $c \in C'$ (conceito da ontologia ou elemento do esquema a serem integrados) é uma *tupla* no formato $c = (T, S)$, onde:

- T é o conjunto de termos (sinônimos) que nomeiam o elemento c. Um termo $t \in T$ é definido como uma relação unária na forma $t(c)$;
- $S = (h, P)$ é a estrutura do elemento c, onde h é a hierarquia (nível em uma seqüência de especializações) na qual o elemento c está localizado, definida como uma relação unária $h(c)$, e $P = (A, R)$ é o conjunto de propriedades do elemento c, onde:
  - A é o conjunto de atributos associados com c. Um atributo $a \in A$ é definido como uma relação $a(c, td)$, sendo td o *tipo de dado* do atributo (uma cadeia de caracteres, um inteiro etc.);

- *R* é o conjunto de relacionamentos de *c* com outros elementos *c' ∈ C* ou *c' ∈ C'*. Um relacionamento *r ∈ R* é definido como uma relação *r*(*c*, *c'*, *cd*), sendo *c'* o conceito relacionado e *cd* a *cardinalidade* do relacionamento *r(c, c')*. Adicionalmente, *r ∈ {g, rt, rc}*, em que *g* é um relacionamento que denota uma geometria; *rt* é um relacionamento topológico, isto é, um tipo especial de relacionamento espacial entre dois conceitos geoespaciais *c* e *c'*; e *rc* é um relacionamento convencional entre dois conceitos *c* e *c'*.

Dada, por exemplo, a estrutura canônica para a ontologia, mais especificamente o conceito *Quarteirao* (Figura 19), pode-se perceber que este possui um conjunto de sinônimos (*Block*, *Quadra*, *Quarteirao*), quatro atributos (*_rgQuarteirao*, *numeroSetor*, *bairro*, *numeroQuarteirao*) e dois relacionamentos (*_ehDelimitado* e *_ehDivididoEm*). Um atributo é formado por seu nome e um tipo de dado (*numeroSetor*, *int*). Um relacionamento é formado pelo nome do relacionamento (*_ehDelimitado*), o outro conceito relacionado (*Logradouro*) e a cardinalidade.

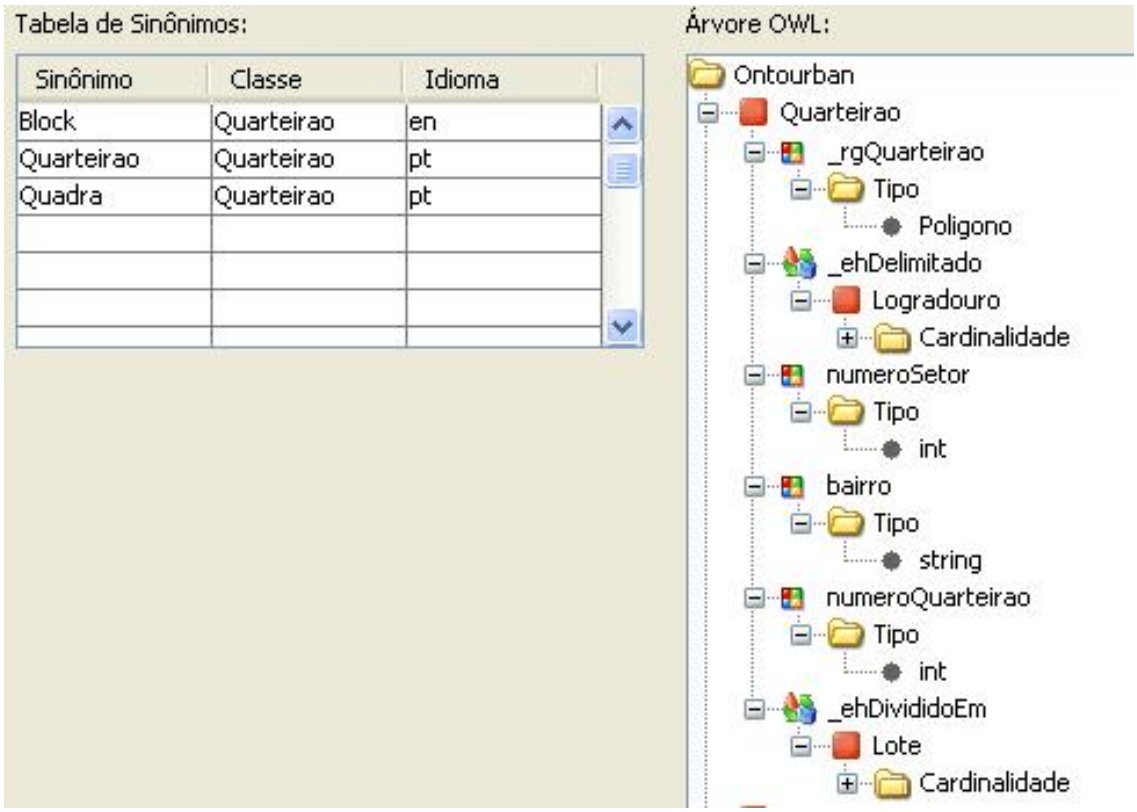


FIGURA 19: Representação do conceito *Quarteirao*, segundo a definição 1

Definição 2: Um conceito geoespacial *cg* é definido como $cg=\{c \in C \mid \exists r(c,c'), r=g\}$, ou seja, *cg* é considerado geoespacial se e somente se tem ao menos um relacionamento *r* do tipo *g*.

Definição 3: Um relacionamento do tipo *rt* é definido como $rt=\{r(c,c') \in R \mid \forall c, c', (c=cg \wedge c'=cg)\}$, ou seja, *rt* pode ocorrer somente entre dois conceitos geoespaciais.

Vale observar ainda que, na GML, um elemento geoespacial se distingue do elemento convencional pela existência de um atributo associado a um tipo de dado geográfico (ponto, linha, polígono). Na OWL (*Web Ontology Language*) não existem tipos de dados geográficos, portanto, este tipo deve ser definido pelo usuário como um novo conceito representando algum atributo ou relacionamento geográfico. A GML ainda não diferencia relacionamentos convencionais e geométricos, exceto no caso de topologias do tipo arco-nó.

As próximas seções detalham cada uma das três partes do método: pré-processamento, determinação de equivalências e catalogação do mapeamento, respectivamente.

4.2 Etapa de Pré-Processamento

A determinação semântica de equivalências entre esquemas GML é uma operação complexa que recebe dois esquemas de entrada e produz um mapeamento entre os elementos desses esquemas que correspondam semanticamente uns aos outros (RAHM e BERNSTEIN, 2001). No contexto deste trabalho, certos elementos de um esquema GML_I são mapeados para certos conceitos na ontologia de domínio, pois se considera que os conceitos na ontologia já estão mapeados para elementos correspondentes em um esquema GML_M, no âmbito do ambiente servidor em que o GML_M está inserido.

Cada elemento é mapeado de acordo com regras de equivalência que especificam como os elementos no GML_I e na ontologia estão relacionados. Para tanto,

considera-se a estrutura e as propriedades dos elementos dos esquemas, como nome, descrição, tipo de dado e tipos de relacionamento (parte-de, é-um etc.).

Porém, como o esquema GML_I e a ontologia são sintaticamente diferentes, isto é, são definidos em linguagens distintas (GML e OWL, respectivamente), é necessário realizar um pré-processamento dos mesmos, a fim de convertê-los para um mesmo padrão de representação, ou seja, para um formato canônico (FROZZA e MELLO, 2007). Esta tarefa é executada por dois módulos, denominados *wrapper* de ontologia e *wrapper* GML.

Como tanto a GML quanto a OWL são linguagens que seguem o padrão XML (*eXtensible Markup Language*), foi adotada uma estrutura em árvore como representação canônica dos dados. Assim, cada elemento GML e cada classe OWL são tratados pelos *wrappers* como um nodo de uma árvore canônica (Figura 20). Esta forma de representação se mostra adequada para representar e manipular dados XML, tornando mais fácil o processamento (DORNELES *et al.* 2004). O entendimento dos dados também é facilitado, a exemplo do que acontece quando se usa estruturas DOM (*Document Object Model*) (DOM, 2007).

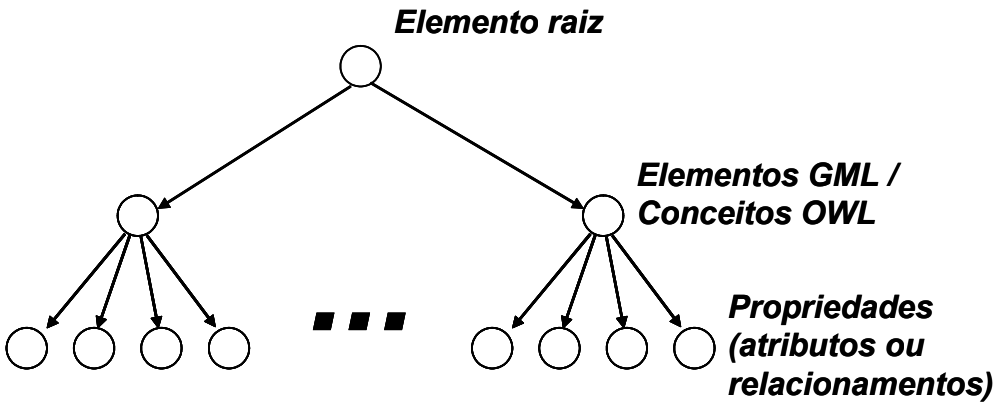


FIGURA 20: Representação de elementos GML e conceitos OWL na forma de árvore canônica

Da mesma forma como acontece com dados XML, cada nodo pode representar um elemento atômico ou complexo. Elementos atômicos contêm valores únicos, como uma cadeia de caracteres, uma data etc. Elementos complexos correspondem a estruturas formadas por outros elementos, atômicos ou complexos.

Nesta estrutura, cada árvore tem três níveis: raiz, elementos complexos e componentes dos elementos complexos (Figura 20). O elemento raiz apenas identifica

se é a árvore da ontologia ou do esquema GML. No nível seguinte, cada elemento complexo representa um conceito da ontologia ou um elemento do esquema GML.

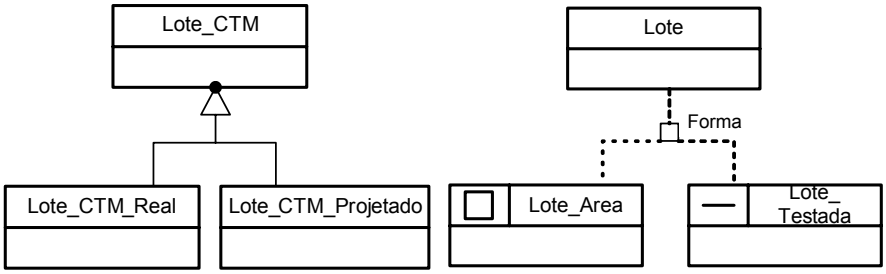


FIGURA 21: Exemplo de especialização e generalização

Na montagem da árvore, optou-se por simplificar o tratamento de hierarquias (especializações e generalizações), permitindo, dessa forma, que o método se torne menos complexo. Tomando como exemplo a Figura 21, em vez de manter as relações de hierarquia, optou-se por criar cada elemento da hierarquia como um elemento independente no segundo nível da árvore. Por exemplo, no caso da especialização de *Lote_CTM* em *Lote_CTM_Real* e *Lote_CTM_Projetado*, foram criados três nodos no segundo nível da árvore, um para cada elemento, sendo que os dois últimos elementos herdam as características (atributos e relacionamentos) de *Lote_CTM*. No cálculo das similaridades, o método informa que todos os três elementos citados são mais ou menos equivalentes a um dado conceito, principalmente em função dos atributos e relacionamentos que eles possuem.

Por fim, no terceiro nível encontram-se os componentes dos elementos complexos, os quais definem as características ou propriedades do elemento. Como visto na seção anterior, essas características são representadas pelos atributos e pelos relacionamentos que cada componente complexo possui. A estrutura de um atributo ou de um relacionamento é tratada pelas métricas de similaridade como um elemento complexo, isto é, um atributo é composto por dois itens, um nome e um tipo de dado, enquanto o relacionamento é composto por três itens, um nome, uma cardinalidade e o nome do elemento com o qual tem relação. Todos esses itens são levados em consideração no cálculo das similaridades.

A criação das duas árvores canônicas permite que o método se torne independente da ontologia e dos esquemas GML, ou seja, no caso de mudanças na forma de criação da ontologia proposta ou da especificação da GML, basta fazer as

devidas adaptações nos *wrappers* que geram as árvores canônicas, preservando o restante do método.

4.2.1 Dicionário de Sinônimos

Na fase de pré-processamento, o *wrapper* de ontologia também gera uma lista de sinônimos associados a cada classe da ontologia, no formato [*SINÔNIMO, CLASSE, IDIOMA*]. Os sinônimos complementam a ontologia, identificando variações conhecidas na denominação de cada termo. Sinônimos podem ser incluídos na ontologia por um especialista do domínio por duas maneiras distintas:

- Manualmente, antes do início do processo de determinação de equivalências, através de uma interface com o usuário própria;
- De forma semi-automatizada, no final do processo, no momento que o especialista interage com o método quando este não consegue definir automaticamente a equivalência de dois termos. Desta forma, a ontologia também pode ser atualizada com novos sinônimos, que são mantidos em uma classe específica.

Conforme pode ser visto no Quadro 7, cada elemento da lista (coluna 1) corresponde a um sinônimo de uma classe (coluna 2) na ontologia. Um sinônimo pode ser o próprio nome da classe da ontologia ou um outro sinônimo qualquer para esta classe. Isto quer dizer que toda classe pode ter pelo menos um sinônimo. Essa abordagem de usar o próprio nome da classe como sinônimo é utilizada por uma questão de simplificação da implementação do método. A indicação do idioma (coluna 3) é utilizada na definição das métricas aplicadas pelo método (ver seção 4.4).

QUADRO 7: Exemplo de lista de sinônimos

SINÔNIMO	CLASSE	IDIOMA
Lote	*Lote*	*pt*
Parcel	*Lote*	*en*
Quadra	*Quarteirao*	*pt*
Quarteirao	*Quarteirao*	*pt*
Block	*Quarteirao*	*en*

A lista de sinônimos é criada a partir das instâncias de objetos de uma classe complementar na ontologia, denominada "*Dicionario*". Especificamente para este método foi adotada essa alternativa de armazenamento dos sinônimos com a intenção de manter todo o conhecimento a respeito do domínio de aplicação centralizado em um

único lugar, ou seja, na ontologia. Entretanto, outras abordagens podem ser estudadas no futuro, como por exemplo, armazenar os sinônimos em um banco de dados relacional.

4.3 Determinação do Grau de Similaridade

Uma vez disponíveis as árvores canônicas representando o esquema GML_I e a ontologia, o próximo passo é a execução do algoritmo de determinação dos graus de similaridade. A base do algoritmo são as regras de equivalência, definidas através de um conjunto de métricas de similaridade.

Como as regras para determinação de equivalências permitem que seja encontrado mais de um candidato válido, é necessário estimar um grau de similaridade através de um valor numérico que pode variar entre 0 e 1 (DORNELES *et al*., 2004). Esse valor é usado pelo método para identificar o melhor candidato a similar automaticamente ou semi-automaticamente. Para tanto, deve ser estabelecido um valor limite (*threshold*) no início do processo, que indica a partir de quando um par de elementos analisados é similar ou não:

- Graus de similaridade abaixo do *threshold* podem ser descartados;
- Se houver um único resultado em que o grau de similaridade for igual ou maior que o *threshold*, então o algoritmo automaticamente o assume como similar;
- Quando mais de um resultado retornar o grau de similaridade igual ou maior que o *threshold*, então o usuário é chamado para intervir no processo. Nestes casos, o usuário pode indicar dentre os resultados qual é o similar e, até mesmo, indicar mais de um resultado.

Na determinação de equivalências, o método leva em consideração o conceito de *granularidade* (RAHM e BERNSTEIN, 2001). Granularidade ao nível de elemento trata de elementos atômicos, ou seja, que contém valores únicos, como pequenas *strings* e valores numéricos. Granularidade ao nível de estrutura trata de elementos complexos, ou seja, estruturas que contêm outros elementos, atômicos ou complexos.

O conceito de granularidade aparece na definição das métricas de similaridade. Dorneles *et al*. (2004) apresentam a seguinte classificação para métricas de similaridade para elementos complexos e elementos atômicos:

- Métricas para Valores Atômicos (MAV – *Metrics for Atomic Values*): são aplicadas a dados simples, como *strings* e números. São dependentes do domínio da aplicação, uma vez que consideram as características dos dados no domínio da aplicação;
- Métricas para Valores Complexos (MCV – *Metrics for Complex Values*): são métricas dependentes de estrutura que podem ser distintamente aplicadas a conjuntos (*tuplas*) ou coleções de valores. No caso deste trabalho, pela natureza dos dados manipulados, somente são encontradas estruturas semelhantes a *tuplas*. Isto se justifica pelo fato de que a definição de um elemento complexo de um esquema GML é composta pela identificação (nome) do elemento e suas propriedades (atributos e relacionamentos), como uma *tupla* em uma tabela relacional.

Esta classificação de métricas proposta por Dorneles *et al*. (2004) apresenta uma taxonomia adequada ao tratamento de dados XML, conforme exposto na seção 4.2.

Além da granularidade, outro ponto importante tratado é a cardinalidade entre os mapeamentos. De acordo com Rahm e Bernstein (2001), um elemento em um esquema pode ser mapeado para um ou mais elementos do outro esquema, possibilitando, assim, relacionamentos de cardinalidade 1:1, 1:*n*, *n*:1 e *n*:*m*. No presente trabalho, as cardinalidades entre mapeamentos são encontradas nas seguintes situações:

a) 1:1 - quando um elemento de um esquema GML equivale a um único conceito na ontologia;
b) 1:*n* - quando um elemento de um esquema GML pode ser equivalente a mais de um conceito na ontologia (por exemplo, no caso de existir a generalização ou a especialização de um conceito);
c) *n*:1 – quando mais de um elemento de um esquema GML pode ser equivalente a um único conceito da ontologia (neste caso, existe a generalização ou a especialização de um elemento GML);

d) *n*:*m* – é uma conseqüência direta da existência de cardinalidades 1:*n* e *n*:1, ou seja, se um conceito da ontologia possui equivalência com vários elementos em GML_I, e vice-versa, então tem-se cardinalidade *n*:*m*.

O fluxo de execução do algoritmo de determinação dos graus de similaridade (Figura 22) tem quatro fases principais:

- Na *primeira fase* é identificado, por meio da lista de sinônimos, a qual conceito da ontologia o nome de um elemento GML pode ser semelhante;
- Na *segunda fase* é calculada a equivalência entre os componentes da estrutura de um elemento GML e os componentes do respectivo conceito na ontologia;
- Na *terceira fase* é calculado o grau de similaridade total entre o elemento GML e o conceito na ontologia;
- Na *quarta* e última fase é feita a catalogação do mapeamento obtido, caso haja equivalência entre o elemento GML e o conceito da ontologia.

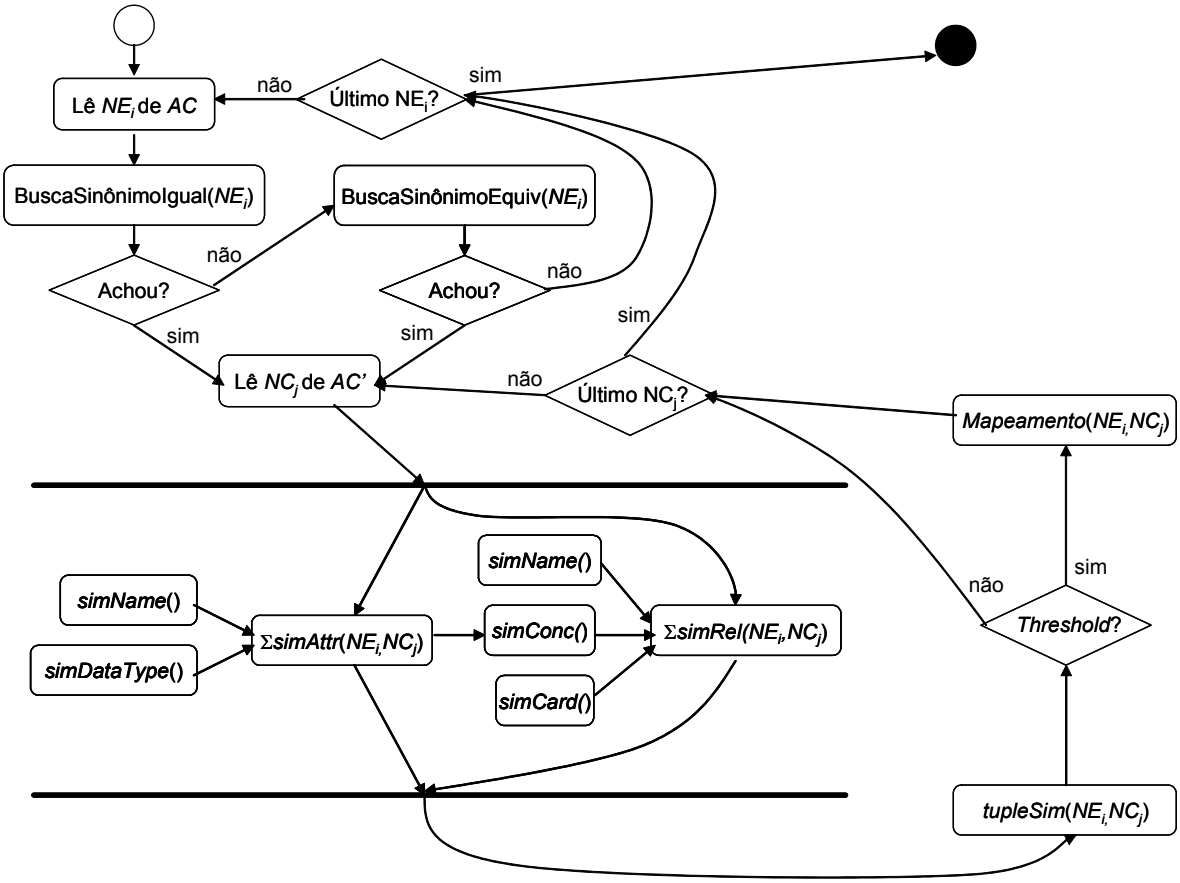


FIGURA 22: Fluxo de execução do algoritmo de determinação de similaridades

A simbologia usada na Figura 22 corresponde a:

- *AC* – corresponde à árvore canônica para um esquema GML;
- NE_i – corresponde a um nodo da árvore *AC*;
- *AC'* – corresponde à árvore canônica da ontologia;
- NC_j – corresponde a um nodo da árvore *AC'*.

A seguir é detalhada cada fase do método. As métricas de similaridade usadas e seus parâmetros são detalhados na seção seguinte.

<u>*Fase 1*</u>:

1. *Lê* NE_i *de AC*: um elemento NE_i da árvore canônica *AC* (GML) é lido;
2. *BuscaSinônimoIgual(*NE_i*)*: O elemento *NEi* é comparado por igualdade contra a lista de sinônimos. Se um sinônimo idêntico for encontrado, passa para o passo 4;
3. *BuscaSinônimoEquiv(*NE_i*)*: O elemento NE_i é comparado por similaridade de nome contra a lista de sinônimos, usando a métrica *simName(*NE_i*, S)*, na qual *S* corresponde à lista de sinônimos. Um sinônimo é dito equivalente se o valor da similaridade for igual ou superior a um *threshold* para nomes (*Tn*), cujo valor padrão foi definido em 0.75 a fim de permitir até 25% de diferenças na comparação das *strings*. Se um sinônimo equivalente não for encontrado, pula para o passo 10 e faz a leitura do próximo NE_i;
4. *Lê* NC_j *de AC'*: o conceito NC_j associado ao sinônimo encontrado para NE_i é lido da árvore canônica *AC'* (ontologia).

<u>*Fase 2*</u>:

5. *Σ simAttr(*NE_i*,* NC_j*)*: calcula a somatória das similaridades dos atributos de NE_i e NC_j. A similaridade entre dois atributos é dada pela soma ponderada das similaridades dos nomes dos atributos (*simName()*) e dos tipos de dados (*simDataType()*). São usados pesos para a similaridade de nomes de atributos (w_{ida}) e tipos de dados (w_{tda}). Se a similaridade dos nomes dos atributos for menor que o *threshold* de nomes (*Tn*), então se considera automaticamente que os dois atributos não são equivalentes;

6. *ΣsimRel(NE_i, NC_j)*: calcula a somatória das similaridades dos relacionamentos de NE_i e NC_j. A similaridade entre dois relacionamentos é dada pela soma ponderada das seguintes similaridades: nomes dos relacionamentos (*simName()*),conceitos relacionados (*simConc()*) e cardinalidades entre os relacionamentos (*simCard()*). Dois relacionamentos são considerados equivalentes se a similaridade dos conceitos relacionados (*simConc()*) for igual ou superior a um *threshold* de relacionamentos (*Tr*), cujo valor padrão é 0.5. Se mais de um par de relacionamentos for identificado como equivalente, é considerado aquele que tiver maior similaridade de nomes.

Fase 3:

7. *tupleSim(NE_i, NC_j)*: calcula-se a similaridade total entre os dois elementos analisados;
8. *Threshold?*: Se a similaridade total for igual ou maior ao valor do *threshold* final (*Tf*), então é feita a catalogação do mapeamento (passo 11). Caso contrário, o par (*NEi*, NC_j) é desconsiderado e parte-se para o próximo par (*NEi*, NC_{j+1})(passo 9);
9. *Último NCj?*: faz a leitura do próximo conceito *NCj* associado a um sinônimo encontrado para NE_i. Se existir, volta ao passo 5;
10. *Último NEi?*: faz a leitura do próximo elemento NE_i. Se existir, volta ao passo 2. Senão, termina o processamento.

Fase 4:

11. *Mapeamento(NE_i, NC_j)*: Faz-se a catalogação do mapeamento obtido (ver seção 4.5). Volta ao passo 9.

4.4 Métricas de Similaridade

O método proposto adapta duas classes de métricas propostas por Dorneles *et al*. (2004) para a determinação dos graus de similaridade semântica entre os elementos do esquema GML importado e da ontologia de domínio:

- Métricas para Valores Complexos (*Metrics for Complex Values* – MCV);

- Métricas para Valores Atômicos (*Metrics for Atomic Values* – MAV).

As métricas do tipo MCV são aplicadas à estrutura dos dados. No caso, como já indicado, pela natureza dos dados tratados, somente são encontradas estruturas semelhantes a *tuplas*.

4.4.1 Métricas para Valores Complexos

Quatro tipos de elementos complexos são identificados neste trabalho, sendo que para cada tipo há uma métrica MCV específica.

- *Nodos NE_i e NC_j*:

Esta métrica foi proposta por Dorneles *et al.* (2004) e adaptada para o contexto deste trabalho. Todas as demais métricas apresentadas foram criadas especificamente para o método proposto. Cada nodo tem um identificador (nome) e diversas propriedades (características) que podem ser definições de atributos ou de relacionamentos. A similaridade dos pares NE_i-NC_j é dada pela equação 1:

$$tupleSim(\varepsilon_p, \varepsilon_d) = \left(w_{nel} * simName(\varepsilon_p, \varepsilon_d)\right) + \left(w_{sar} * \frac{\sum (sim(\varepsilon_p^i, \varepsilon_d^j))}{\max(m,n)} \right) \quad \text{Eq. 1}$$

onde:

- p: conjunto de nodos NE_i da árvore canônica do esquema GML_I;
- d: conjunto de nodos NC_j da árvore canônica da ontologia;
- ε_p: um nodo (NE_i) do conjunto p;
- ε_d: um nodo (NC_j) do conjunto d;
- ε_p^i : uma propriedade de ε_p (atributo ou relacionamento);
- ε_d^j : uma propriedade de ε_d (atributo ou relacionamento);
- *simName()*: é a métrica para cálculo da similaridade de nomes;
- w_{nel}: é o peso que a similaridade dos nomes tem na métrica *tupleSim()*;
- w_{sar}: é o peso que a similaridade das propriedades dos dois nodos tem na métrica *tupleSim()*. A somatória de w_{nel} e w_{sar} deve ser igual a 1;
- *sim()*: é a similaridade das propriedades dos nodos;
- n e m: número de nodos filhos de ε_p e ε_d, respectivamente, ou seja, a quantidade de atributos e relacionamentos em cada nodo;

- *max()*: o maior valor entre *m* e *n*.

Na métrica *tupleSim()*, cada nodo filho ε_p^i de ε_p é comparado contra todos os nodos filhos ε_d^j de ε_d, até identificar um com nome equivalente e estrutura semelhante (nome e tipo de dado, no caso de um atributo; nome, conceito relacionado e cardinalidade, no caso de um relacionamento). As comparações entre nodos filhos são feitas por métricas específicas para atributos (*simAttr()*) e relacionamentos (*simRel()*). Através das árvores canônicas é possível identificar se um nodo filho representa um atributo ou um relacionamento. A função *max()* retorna o maior número de nodos filho entre ε_p e ε_d.

A métrica *tupleSim()* é executada em duas etapas:

- Na primeira etapa, calcula-se apenas a similaridade dos nomes dos nodos NE_i e NC_j, além da somatória das similaridades dos nodos filhos que correspondam a atributos de ε_p e ε_d, para os quais é aplicada a métrica *simAttr()*;
- Na segunda etapa, são calculadas a somatória das similaridades para os nodos filhos que correspondam aos relacionamentos, usando a métrica *simRel()*.

Essa seqüência de execução é necessária para evitar situações de *loop* infinito que podem ocorrer devido às dependências entre os relacionamentos, conforme explicado a seguir.

- *Definições de atributos simples*:

Atributos simples são compostos por um identificador (nome) e um tipo de dado (*string*, *integer*, *polygon* etc.). A equação 2 representa a métrica usada para definir a similaridade de atributos.

$$simAttr(\varepsilon_p^i, \varepsilon_d^j) = w_{ida} * simName(\varepsilon_p^i, \varepsilon_d^j) + w_{tda} * simDataType(\varepsilon_p^i, \varepsilon_d^j) \quad \text{Eq. 2}$$

onde:

- ε_p^i : nodo filho de um elemento da árvore GML;
- ε_d^j : nodo filho de um elemento da árvore OWL;

- w_{ida}: representa o peso (*weight*) que a similaridade de nomes tem no contexto da similaridade de atributos;
- *simName()*: similaridade dos nomes dos atributos;
- w_{tda}: representa o peso (*weight*) que a similaridade dos tipos de dados tem no contexto da similaridade de atributos;
- *simDataType()*: similaridade dos tipos de dados.

Os pesos para a similaridade de nomes de atributos e de tipos de dados (w_{ida} e w_{tda}) representam a importância do nome e do tipo de dado no contexto semântico tratado. O valor destes dois pesos é definido por meio de parâmetros no início do processo e a somatória de ambos deve ser igual a 1. O valor padrão é definido como {w_{ida}=0.5, w_{tda}=0.5}, o que estabelece igual importância para os dois parâmetros.

- *Definições de relacionamentos*:

Relacionamentos são definidos pelo nome do relacionamento, o conceito relacionado e a cardinalidade. A equação 3 representa a métrica para definir a similaridade de relacionamentos.

$$simRel(\varepsilon_p^i, \varepsilon_d^j) = w_{idr} * simName(\varepsilon_p^i, \varepsilon_d^j) + w_{cor} * simConc(\varepsilon_p^i, \varepsilon_d^j) + w_{car} * simCard(\varepsilon_p^i, \varepsilon_d^j) \quad \text{Eq. 3}$$

onde:

- ε_p^i : nodo filho de um elemento da árvore GML;
- ε_d^j : nodo filho de um elemento da árvore OWL;
- w_{idr}: representa o peso (*weight*) que a similaridade de nomes tem no contexto da similaridade de relacionamentos;
- *simName()*: similaridade dos nomes dos relacionamentos;
- w_{cor}: representa o peso (*weight*) da similaridade de conceitos;
- *simConc()*: similaridade dos conceitos referenciados pelos relacionamentos;
- w_{car}: representa o peso (*weight*) da similaridade de cardinalidades;
- *simCard()*: similaridade das cardinalidades.

Da mesma forma que ocorre com a similaridade de atributos, os pesos para a similaridade de nomes, conceitos relacionados e cardinalidades (w_{idr}, w_{cor} e w_{car},

respectivamente) representam a importância dos nomes, conceitos e tipos de dados no contexto semântico do relacionamento. O valor destes três pesos é definido por meio de parâmetros no início do processo e a somatória deles ser 1. O valor padrão é estabelecido, respectivamente, em {w_{idr}=0.3, w_{cor}=0.5, w_{car}=0.2}.

- *Similaridade de conceitos*:

Dado um relacionamento r = (n, c', cd), em que c' corresponde ao conceito relacionado, a métrica de similaridade de conceitos (*simConc()*) procura identificar se um conceito c' presente em um relacionamento de um elemento GML é equivalente a um conceito c' presente em um relacionamento na classe da ontologia associada a este elemento. A equação 4 representa a métrica de similaridade de conceitos aplicada no método.

$$simConc(\varepsilon_p^i, \varepsilon_d^j) = \left(w_{nel} * simName(conc_1, conc_2)\right) + \left(w_{sar} * \frac{\sum simAttr(conc_1, conc_2)}{max(n, m)}\right) \quad \text{Eq. 4}$$

onde:

- ε_p^i : nodo filho de um elemento da árvore GML;
- ε_d^j : nodo filho de um elemento da árvore OWL;
- $conc_1$: conceito c' do relacionamento no elemento GML;
- $conc_2$: conceito c' do relacionamento na classe OWL;
- w_{nel}: peso da similaridade dos nomes dos conceitos c';
- w_{sar}: peso da similaridade dos atributos dos conceitos c';
- *simName*: similaridade dos nomes dos conceitos;
- $\Sigma simAttr(conc_1, conc_2)$: somatória das similaridades dos atributos de ε_p^i e ε_d^j;
- n e m: quantidade de atributos de ε_p^i e ε_d^j, respectivamente;
- *max()*: maior valor entre n e m.

A similaridade de dois conceitos é dada em função da similaridade de seus nomes mais a similaridade dos atributos. Neste caso, não são considerados os relacionamentos existentes nos conceitos c', visto que isso poderia gerar situações de

loop infinito em função das dependências existentes nos relacionamentos. Por exemplo, supondo *simConc(A, B)*, no momento em que a similaridade de *A* está sendo processada e é identificado um relacionamento com *B*. Então, o processamento de *A* é interrompido para que se inicie o processamento de *B*. Nesta situação, pode ocorrer um *loop* infinito se for identificado um relacionamento de *B* com *A*, gerando assim um ciclo recursivo.

Para compensar a falta dos relacionamentos em *c'*, dois conceitos são ditos equivalentes se o resultado de *simConc()* for igual ou superior a um *threshold* de conceitos (*Tr*), cujo valor padrão é definido em 0.5.

4.4.2 Métricas para Valores Atômicos

As métricas do tipo MAV são usadas para calcular a similaridade de dados simples, como *strings*, datas e números. Elas são dependentes do domínio da aplicação, já que consideram as características dos dados da aplicação. Neste trabalho são identificados três tipos de dados simples: nomes, tipos de dados e cardinalidades.

- *Similaridade de nomes*:

Um caso comum de elemento atômico são os nomes usados para identificar elementos, atributos e relacionamentos, os quais são do tipo *string*. A determinação das equivalências de nomenclaturas usa uma abordagem lingüística, mais precisamente, ela verifica a compatibilidade de textos (RAHM e BERNSTEIN, 2001). Esta abordagem determina a equivalência entre elementos de um esquema por meio da igualdade ou similaridade do texto.

A similaridade de textos pode ser definida e medida através de várias abordagens, como (RAHM e BERNSTEIN, 2001):

- *Igualdade de nomes*: assegura que nomes idênticos realmente sustentam a mesma semântica;
- *Igualdade de representação canônica do texto depois de algum pré-processamento*: útil para derivar prefixos ou sufixos (por exemplo: *CName = customer name*);
- *Igualdade de sinônimos*;
- *Igualdade de hiperônimos*: requer o uso de *thesaurus* ou dicionários;

- *Similaridade de textos baseados em substrings comuns, distância de edição, pronúncia, sonoridade etc.*: aplica métricas de similaridade;
- *Equivalências de textos pré-determinadas pelo usuário*.

Estas abordagens podem ser usadas isoladamente no algoritmo de equivalência ou combinadas, a fim de prover alternativas quando uma falhar. Por exemplo, na primeira fase do método é usada uma abordagem de *igualdade de sinônimos* para identificar se um elemento NE_i existe na lista de sinônimos tal qual como foi escrito (seção 4.3, passo 2). Caso não seja encontrado um sinônimo idêntico, então se aplica uma métrica de *similaridade de textos* (seção 4.3, passo 3).

Algumas métricas de similaridade de *strings* são encontradas na literatura (CHAPMAN, 2006), como as métricas *Jaro Winkler*, distância de *Levenshtein* e distância de *Hamming*. No método proposto, é possível selecionar a métrica a ser utilizada para definir a similaridade de nomes, usando como parâmetro o idioma obtido na lista de sinônimos. No momento da definição dos parâmetros iniciais, pode ser indicada qual das métricas suportadas é aplicada em cada idioma. Assim, o método se torna mais flexível e fica livre para escolher a métrica mais adequada em cada situação.

Essa abordagem é usada, uma vez que certas métricas produzem melhores resultados em um idioma específico, em função da forma como as palavras são construídas. Por exemplo, atualmente o método suporta a métrica *Jaro Winkler* (que valoriza a existência de prefixos) para palavras na língua portuguesa, pois os nomes de elementos podem ser compostos por mais de uma palavra, como *LoteCTM* e *LoteReal*. Já, para a língua inglesa, o método usa a distância de *Levenshtein* pois, nesse caso, a construção das palavras não mantém a mesma estrutura de prefixos que acontece na língua portuguesa (*MTRParcel* e *ActualParcel*). Para outras línguas, é utilizado o maior valor retornado pelas duas métricas acima. A equação 5 descreve a métrica de similaridade de nomes aplicada no método.

$$simName(\varepsilon_p^i, \varepsilon_d^j) = sim(str_1, str_2) \qquad \text{Eq. 5}$$

onde:

- ε_p^i : nodo filho de um elemento da árvore GML;
- ε_d^j : nodo filho de um elemento da árvore OWL;

- str_1 e str_2: representam as duas *strings* sendo analisadas;
- *sim()*: representa uma métrica de similaridade de *strings* escolhida dinamicamente, conforme o idioma dos termos que estão sendo analisados.

A métrica *simName()* é aplicada em três situações distintas no algoritmo para identificação de similaridades entre:

- os identificadores nas árvores canônicas da OWL e da GML;
- os identificadores de atributos;
- os identificadores de relacionamentos.

Nos dois primeiros casos, é dito que dois dados são similares se o valor retornado pela métrica *simName*() for igual ou superior a um *threshold* de nomes (*Tn*), cujo valor padrão é 0.75.

- *Similaridade de tipos de dados*:

A métrica de similaridade de tipos de dados (*simDataType()*) procura identificar se os tipos de dados de dois atributos distintos são equivalentes. Essa equivalência é dada por um mapeamento prévio dos tipos de dados presentes em um esquema GML em relação aos tipos de dados usados na ontologia. Há dois conjuntos de tipos de dados para definição de atributos sendo analisados: *tipos de dados convencionais* e *tipos de dados geométricos*.

Quanto aos tipos de dados convencionais, a OWL permite usar a maioria dos tipos de dados disponíveis para XML *Schema*. A especificação da OWL (OWL, 2007) recomenda o uso de 35 tipos de dados diferentes, porém, a ferramenta *Protégé* (PROTÉGÉ, 2007), usada para criação da ontologia deste trabalho, oferece ao usuário apenas oito tipos de dados: *time*, *dateTime*, *date*, *string*, *int*, *float*, *boolean* e *any*. Este último pode assumir qualquer tipo definido para a XML *Schema*.

A GML também faz uso dos tipos definidos internamente para a XML *Schema*, uma vez que é uma gramática codificada segundo suas regras. Na especificação da GML 3.1.1 (OGC, 2004), por exemplo, são usados pelo menos treze tipos de dados diferentes: *string*, *integer*, *double*, *positiveInteger*, *decimal*, *nonNegativeInteger*, *boolean*, *date*, *dateTime*, *time*, *duration*, *gYearMonth* e *gYear*. Os tipos de dados convencionais em XML *Schema* podem ser classificados em (XML SCHEMA, 2004):

- *Tipos de dados primitivos*: são aqueles que não são definidos em termos de outros tipos de dados (existem 19 tipos primitivos);
- *Tipos de dados derivados*: são aqueles que são definidos em termos de outros tipos de dados (existem 25 tipos derivados).

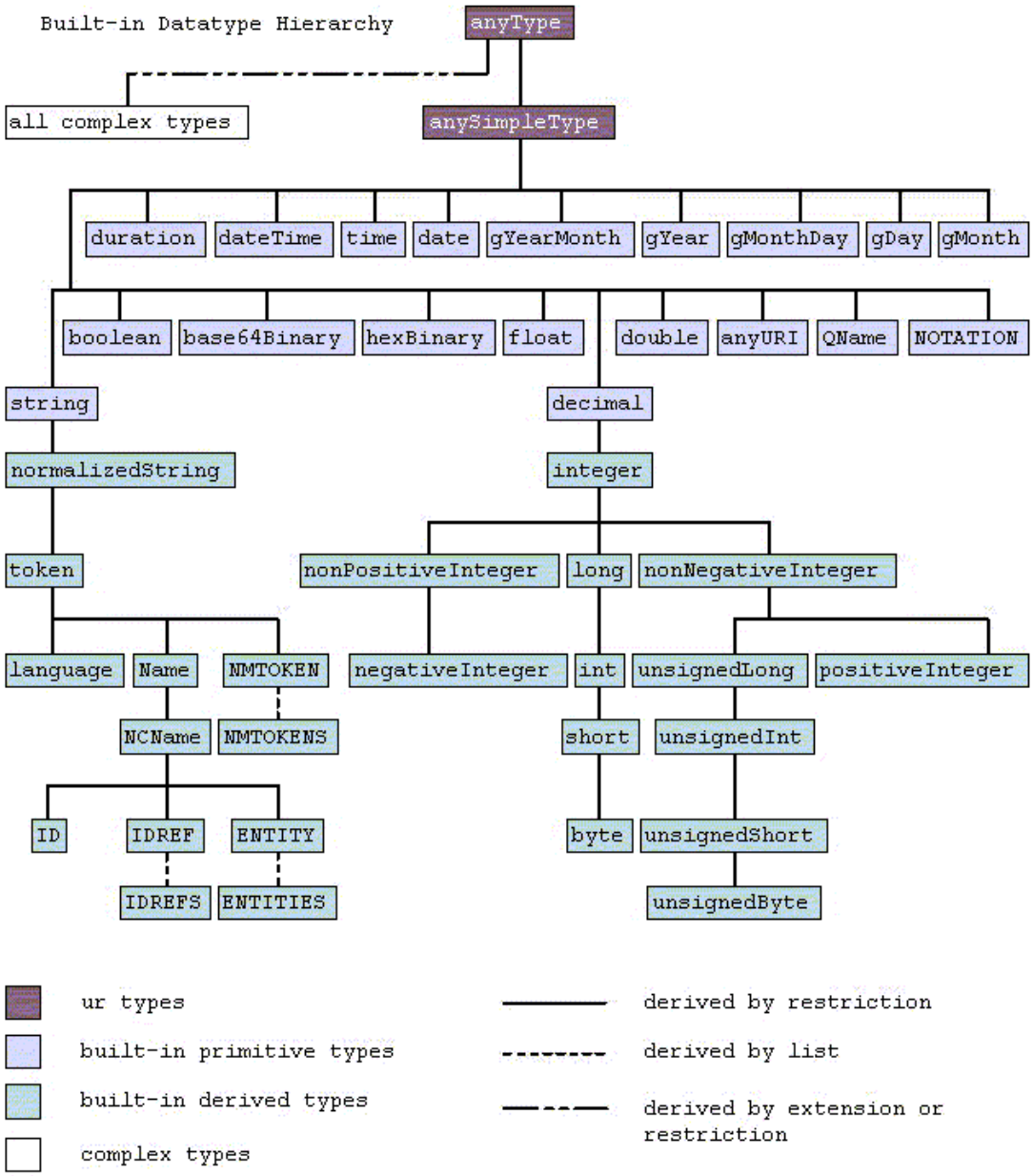


FIGURA 23: Hierarquia dos tipos de dados internos da XML *Schema*
(FONTE: XML SCHEMA, 2004)

A Figura 23 apresenta a hierarquia dos tipos de dados primitivos e derivados definidos para a XML *Schema* (XML SCHEMA, 2004). Como nem todos os tipos de dados existentes na XML *Schema* são usados para definir atributos convencionais, o

Quadro 8 apresenta uma proposta de mapeamento para este trabalho, a qual é utilizada para retornar os resultados na métrica *simDataType()*.

QUADRO 8: Mapeamento entre tipos de dados convencionais da OWL em relação à GML

OWL	GML	Grau de Similaridade	Representação do tipo em GML
• Tipos de dados para cadeia de caracteres			
String	*string*	1.0	Cadeia de caracteres
• Tipos de dados para numerais			
Int	*int*	1.0	{-2147483648, ..., 2147483647}
	integer	0.9	{..., -2, -1, 0, 1, 2, ...}
	positiveInteger	0.5	{1, 2, ...}
	nonNegativeInteger	0.6	{0, 1, 2, ...}
	string	0.4	Cadeia de caracteres
Float	*float*	1.0	*m**2^*e*; *m* => *integer* <=2^24; *e* = {-149, ..., 104}
	decimal	0.5	*i**10^-*n*; *i*,*n* => *integer*; *n*>=0
	double	0.5	*m**2^*e*; *m* => *integer* <=2^53; *e* = {-1075, ..., 970}
	string	0.4	Cadeia de caracteres
• Tipos de dados para datas e horas			
dateTime	*dateTime*	1.0	2002-10-10T12:00:00
	duration	0.2	P*n*Y*n* M*n*DT*n*H *n*M*n*S
	time	0.5	hh:mm:ss.sss
	date	0.5	yyyy-mm-ddT00:00:00
	gYearMonth	0.2	CCYY-MM
	gYear	0.2	CCYY
	string	0.4	Cadeia de caracteres
date	*date*	1.0	yyyy-mm-ddT00:00:00
	duration	0.2	P*n*Y*n* M*n*DT*n*H *n*M*n*S
	time	0.0	hh:mm:ss.sss
	dateTime	0.5	2002-10-10T12:00:00
	gYearMonth	0.2	CCYY-MM
	gYear	0.2	CCYY
	string	0.4	Cadeia de caracteres
time	*time*	1.0	hh:mm:ss.sss
	duration	0.2	P*n*Y*n* M*n*DT*n*H *n*M*n*S
	date	0.0	yyyy-mm-ddT00:00:00
	dateTime	0.5	2002-10-10T12:00:00
	gYearMonth	0.0	CCYY-MM
	gYear	0.0	CCYY
	string	0.4	Cadeia de caracteres
• Tipos de dados binários			
Boolean	*boolean*	1.0	{true, false, 1, 0}
	string	0.5	Cadeia de caracteres

Assumindo que a ontologia representa a base de conhecimento do método, foram considerados apenas os principais tipos de dados que podem ser mapeados para tipos equivalentes na ontologia e desconsiderados os limites máximos e mínimos dos tipos de dados usados para representar números no esquema GML. O grau de similaridade é definido arbitrariamente, uma vez que não é possível estimar pela

comparação da faixa de valores permitidos, pois alguns tipos de dados tendem a um conjunto infinito. Em algumas situações, a semântica do tipo de dado foi observada, como por exemplo, no caso dos tipos de dados para representar datas e horas.

Quanto aos tipos de dados geográficos, também foi definido o mapeamento (Quadro 9). Diferentemente dos tipos de dados convencionais, neste caso há um problema relacionado com a ontologia, uma vez que a OWL não possui uma definição específica para dados geográficos. Por essa razão criou-se na ontologia uma classe específica, com especialização nos três tipos básicos de dados geográficos: polígono, linha e ponto. O mapeamento, então, é feito entre os tipos de dados geográficos presentes na GML com estas respectivas especializações.

QUADRO 9: Mapeamento entre tipos de dados convencionais da OWL em relação à GML

Classe OWL	**GML**	**Similaridade**
Ponto	*gml:Point*	1.0
Linha	*gml:Curve*	1.0
	gml:LineString	
	gml:OrientableCurve	
	gml:CompositeCurve	
Polígono	*gml:Solid*	1.0
	gml:CompositeSolid	
	gml:Polygon	

(FONTE: adaptado de HESS, 2004)

A equação 6 representa a métrica de similaridade de tipos de dados.

$$simDataType(\varepsilon_p^i, \varepsilon_d^j) = sim(dt_1, dt_2) \qquad \text{Eq. 6}$$

onde:

- ε_p^i : nodo filho de um elemento da árvore GML;
- ε_d^j : nodo filho de um elemento da árvore OWL;
- dt_1 e dt_2: representam os dois tipos de dados sendo analisados;
- *sim()*: retorna um valor conforme os Quadros 8 e 9.

- *Similaridade de cardinalidades*:

Dada a definição de um relacionamento r = (n, c', cd), a métrica de similaridade de cardinalidades (*simCard()*) procura identificar se as cardinalidades *cd* existentes em dois relacionamentos avaliados são equivalentes. Para tanto, ela usa a mesma abordagem da métrica para similaridades entre tipos de dados, ou seja, é definido um mapeamento entre cardinalidades, conforme apresentado no Quadro 10.

Cardinalidades mínimas e opcionais não foram tratadas, uma vez que não são usadas no cálculo da equivalência dos esquemas porque a ontologia, como foi definida, apenas indica se um relacionamento tem uma única instância (*single*) ou tem mais de uma instância (*multiple*).

QUADRO 10: Mapeamento entre cardinalidades das classes OWL em relação aos elementos GML

Classe OWL	Elemento GML	Similaridade
single	*maxOccurs = 1*	1.0
single	*maxOccurs > 1*	0.5
multiple	*maxOccurs = 1*	1.0
multiple	*maxOccurs > 1*	1.0

Na ontologia, propriedades do tipo objeto são usadas para representar o relacionamento entre dois conceitos. A cardinalidade deste relacionamento pode ser definida de duas formas: relacionamento funcional (*single* - aceita um único valor) ou não-funcional (*multiple* - aceita múltiplos valores). Além disso, os relacionamentos são direcionados e definidos separadamente, ou seja, o relacionamento de *c* para *c'* cria uma propriedade do tipo objeto em *c*, enquanto que o relacionamento de *c'* para *c* (inverso do relacionamento) cria uma propriedade do tipo objeto em *c'*.

```
01  <complexType name="LoteAreaType">
02     <complexContent>
03        <extension base="urb:LoteType">
04           <sequence>
                ...
05              <element name="possuiDivisaLote"
                  type="urb:DivisaLotePropertyType" minOccurs="0"
                  maxOccurs="unbounded">
                  ...
06              </element>
                ...
07              <element name="existeCemiterio"
                  type="urb:CemiterioPropertyType" minOccurs="0">
                  ...
08              </element>
                ...
09           </sequence>
10        </extension>
11     </complexContent>
12  </complexType>
13  <complexType name="LoteAreaPropertyType">
14     <sequence>
15        <element ref="urb:LoteArea" minOccurs="0"/>
16     </sequence>
17     <attributeGroup ref="gml:AssociationAttributeGroup"/>
18  </complexType>
```

FIGURA 24: Exemplo de definição de relacionamento em um esquema GML

No esquema GML, um relacionamento é definido por meio de um *complexType*, cujas propriedades correspondem a referências aos elementos relacionados (ver destaque na Figura 24). As cardinalidades são definidas por meio dos

atributos *minOccurs* e *maxOccurs* em cada referência aos elementos relacionados (linhas 05 e 07 na Figura 24). Quando um dos dois atributos não é indicado, assume-se o valor padrão (1), enquanto que o valor *unbounded* significa "sem limites".

A equação 7 representa a métrica de similaridade de cardinalidades.

$$simCard(\varepsilon_p^i, \varepsilon_d^j) = sim(card_1, card_2) \quad \text{Eq. 7}$$

onde:

- ε_p^i : nodo filho de um elemento da árvore GML;
- ε_d^j : nodo filho de um elemento da árvore OWL;
- $card_1$ e $card_2$: representam as cardinalidades sendo analisadas;
- *sim()*: retorna um valor conforme o Quadro 10.

4.5 Mapeamento das Equivalências

Conforme pode ser visto na Figura 22, ao final de cada ciclo de processamento do método, o valor obtido na métrica *tupleSim()* é comparado contra um *threshold* final (*Tf*). Se o resultado da métrica for igual ou superior ao *Tf*, então o par de elementos analisados é considerado equivalente e é encaminhado para o módulo de mapeamento. O valor padrão para este *threshold* é definido em 0.75, levando em conta que dois elementos podem ser até 25% diferentes para serem considerados equivalentes.

Neste momento duas situações podem acontecer:

- É encontrado um único conceito da ontologia equivalente ao elemento GML_I;
- Mais de um conceito da ontologia é definido como equivalente ao elemento GML_I.

No primeiro caso, é feito o mapeamento automático do elemento GML_I para a ontologia e, possivelmente, para um elemento GML_M que já possui mapeamento para a ontologia. No segundo caso, o mapeamento é feito de forma semi-automática, ou seja, o método apresenta ao especialista todas as equivalências encontradas e este determina quais estão corretas. Assim, o especialista pode tomar uma de quatro decisões possíveis:

- Nenhuma equivalência está correta;

- Todas as equivalências estão corretas: ocorre quando há especializações de um conceito OWL;
- Somente uma equivalência é correta;
- Mais de uma equivalência é correta: no caso de existirem especializações de um conceito OWL e somente parte dessas especializações são consideradas equivalentes.

A partir dessas decisões pode-se, também, iniciar um processo de atualização da ontologia, incluindo-se novos sinônimos, atributos e relacionamentos a um conceito já existente ou, até mesmo, incluir um novo conceito na ontologia. Este processo de atualização da ontologia está fora do escopo deste trabalho.

Uma vez confirmadas as equivalências semânticas entre o esquema GML importado e a ontologia, a ferramenta cataloga, em tabelas relacionais de mapeamento, quais elementos do GML_I correspondem semanticamente a elementos do GML_M.

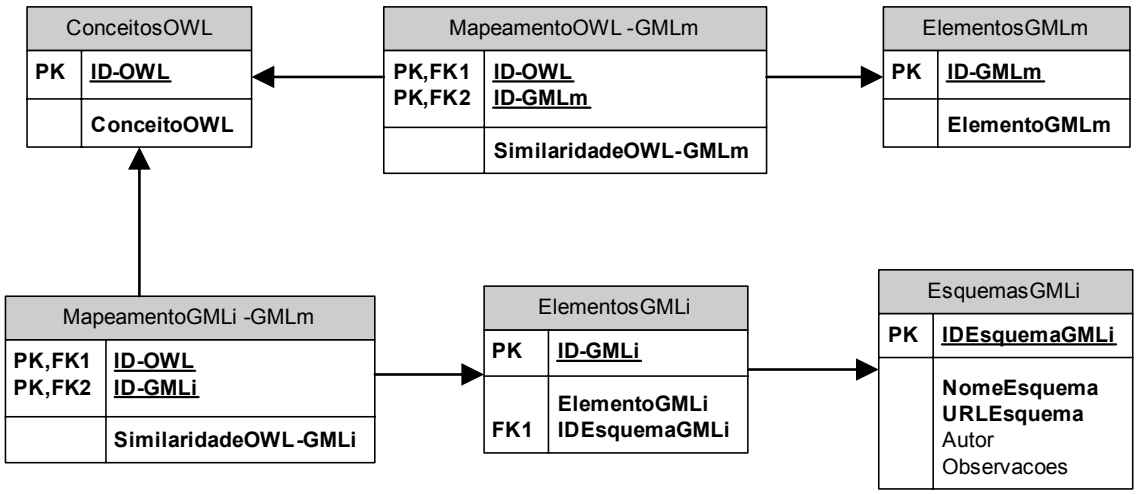


FIGURA 25: Esquema relacional para o mapeamento entre esquemas GML

A Figura 25 apresenta o esquema relacional utilizado para os mapeamentos. Neste esquema são definidas seis tabelas:

- *ConceitosOWL* – mantém um registro para cada conceito presente na ontologia;
- *ElementosGMLm* – mantém um registro para cada elemento do esquema GML_M que possui equivalência com um conceito na ontologia;
- *MapeamentoOWL-GMLm* – mantém os mapeamentos entre conceitos presentes na ontologia e elementos do esquema GML_M. Um elemento

GML_M pode ser associado a mais de um conceito na ontologia. Isso ocorre quando há especializações dos conceitos da ontologia;

- *EsquemasGMLi* – mantém metadados dos esquemas importados (Nome do esquema, URL do esquema, Autor e Observações);
- *ElementosGMLi* – mantém um registro para cada elemento do esquema GML_I que possui equivalência com um conceito na ontologia;
- *MapeamentoGMLi-GMLm* – mantém os mapeamentos entre os elementos do esquema GML_I e os conceitos na ontologia e, por conseqüência, com os mapeamentos para os elementos do esquema GML_M.

A primeira execução do método deve ser realizada tendo como entrada a ontologia e o esquema GML_M. Nesta etapa, é feita a calibragem do sistema, ou seja, identificados os melhores valores para pesos e *thresholds* (limiares), levando em consideração os conceitos da ontologia e os elementos presentes no esquema GML principal (GML_M). A calibragem deve ser feita em um processo de tentativa e erro, até alcançar um mapeamento aceitável para o esquema GML_M. Como resultado final, são preenchidas as tabelas *ConceitosOWL*, *ElementosGMLm* e *MapeamentoOWL-GMLm*.

Para a calibragem, os seguintes valores são definidos:

- w_{nel} – peso da similaridade de nomes dos elementos (valor padrão = 0.5);
- w_{sar} – peso da similaridade de atributos dos elementos (valor padrão = 0.5);
- w_{idr} – peso da similaridade de nomes nos relacionamentos dos elementos (valor padrão = 0.35);
- w_{cor} – peso da similaridade de conceitos relacionados nos relacionamentos dos elementos (valor padrão = 0.35);
- w_{car} – peso da similaridade da cardinalidade em relacionamentos dos elementos (valor padrão = 0.3);
- w_{ida} – peso da similaridade do nome de atributos dos elementos (valor padrão = 0.5);
- w_{tda} – peso da similaridade de tipo de dados dos atributos dos elementos (valor padrão = 0.5);
- *Tn – threshold* para nomes (valor padrão = 0.75);
- *Tr – threshold* para conceitos em relacionamentos (valor padrão = 0.5);

- *Tf – threshold* final (valor padrão = 0.75).

Os valores padrões acima definidos foram obtidos por meio de testes realizados com a ontologia e o esquema GML_M, nos quais se procurou atribuir pesos que refletissem o mesmo grau de importância para cada componente da estrutura de um atributo e de um relacionamento, enquanto que os *thresholds* foram utilizados como filtro para determinar as equivalências corretas. A definição destes valores é dependente da ontologia e do esquema GML principal existentes no ambiente em que o método é executado. Uma vez definidos estes parâmetros, os mesmos não podem mais ser alterados para as próximas execuções do método, nas quais são utilizados os esquemas importados (GML_I).

Uma vez feito o mapeamento do GML_M com a ontologia, pode-se definir o mapeamento dos esquemas GML_I. Informações gerais sobre cada esquema GML_I são encontradas na tabela *EsquemasGMLi* e cada elemento equivalente encontrado em GML_I gera um registro nas tabelas *ElementosGMLi* e *MapeamentoGMLi-GMLm*.

4.6 Conclusão

Este capítulo apresentou o método para determinação de equivalências semânticas entre esquemas GML, objetivo principal do trabalho. A partir de definições gerais dos componentes de um elemento GML ou OWL, o método é detalhado em três etapas: pré-processamento, determinação dos graus de similaridade e mapeamento das equivalências.

Na etapa de pré-processamento, *wrappers* traduzem os esquemas GML e a ontologia para um formato canônico, que facilita o processamento na etapa seguinte.

A etapa de determinação dos graus de similaridade é a parte principal do método. Ela é caracterizada pelo algoritmo que orienta o fluxo de execução do método e pelas métricas de similaridade. Ao todo, foram definidas sete métricas de similaridade, sendo quatro para elementos complexos e três para elementos atômicos.

A similaridade semântica entre dois elementos, objetivo principal do trabalho, é obtida por meio da aplicação destas métricas em conjunto. Elas levam em consideração a nomenclatura e a estrutura dos elementos e seus componentes. De

acordo com a função de cada métrica, são definidos pesos que determinam maior ou menor importância semântica para o resultado da métrica. Ainda, existe também um conjunto de *thresholds* que serve para filtrar as equivalências corretas.

Ao final, os elementos que tiverem suas equivalências corretamente definidas são mapeados e armazenados em tabelas relacionais de um banco de dados. O mapeamento pode ser automático, quando for identificada apenas uma equivalência, ou semi-automático, quando forem identificadas mais de uma equivalência para o mesmo elemento, caso em que o especialista é solicitado para interagir com o método.

A simplicidade foi uma das características que nortearam o desenvolvimento do método, ou seja, na medida do possível, procurou-se manter a simplicidade em relação à definição de conceitos e métricas de similaridade, a fim de que essa simplicidade também se refletisse no algoritmo e na implementação do método.

Por outro lado, o método foi construído com base em um pequeno conjunto de conceitos (*quadras* e *lotes*) que fazem parte do domínio do cadastro urbano e, conseqüentemente, quando este conjunto de conceitos for ampliado, algumas modificações ainda não previstas podem ser necessárias.

Por fim, com relação aos esquemas GML, não foi considerado o uso de nomes voltados para a implementação, ou seja, casos em que os identificadores em um esquema GML são gerados a partir de identificadores existentes em um banco de dados. Esses identificadores, muitas vezes, são códigos que abreviam o nome do elemento que ele representa, por exemplo, "*número do logradouro*" e "*numlog*". Para estes casos, posteriormente deve ser estudada uma alternativa para compatibilização de nomes.

5 ESTUDO DE CASO

Uma ferramenta que dá suporte ao método proposto no capítulo 4 foi implementada (ROSA, 2006). Este capítulo descreve a interface da ferramenta e, em seguida, apresenta um estudo de caso realizado com a sua utilização. Descreve os dados de entrada usados nos testes e demonstra passo-a-passo a aplicação do método para um elemento de um esquema GML. Ao final, analisa os resultados de alguns experimentos realizados com a ferramenta, através das métricas *recall* e *precision*.

5.1 Interface da Ferramenta

A Figura 26 apresenta a interface principal da ferramenta de determinação de equivalências semânticas entre esquemas GML.

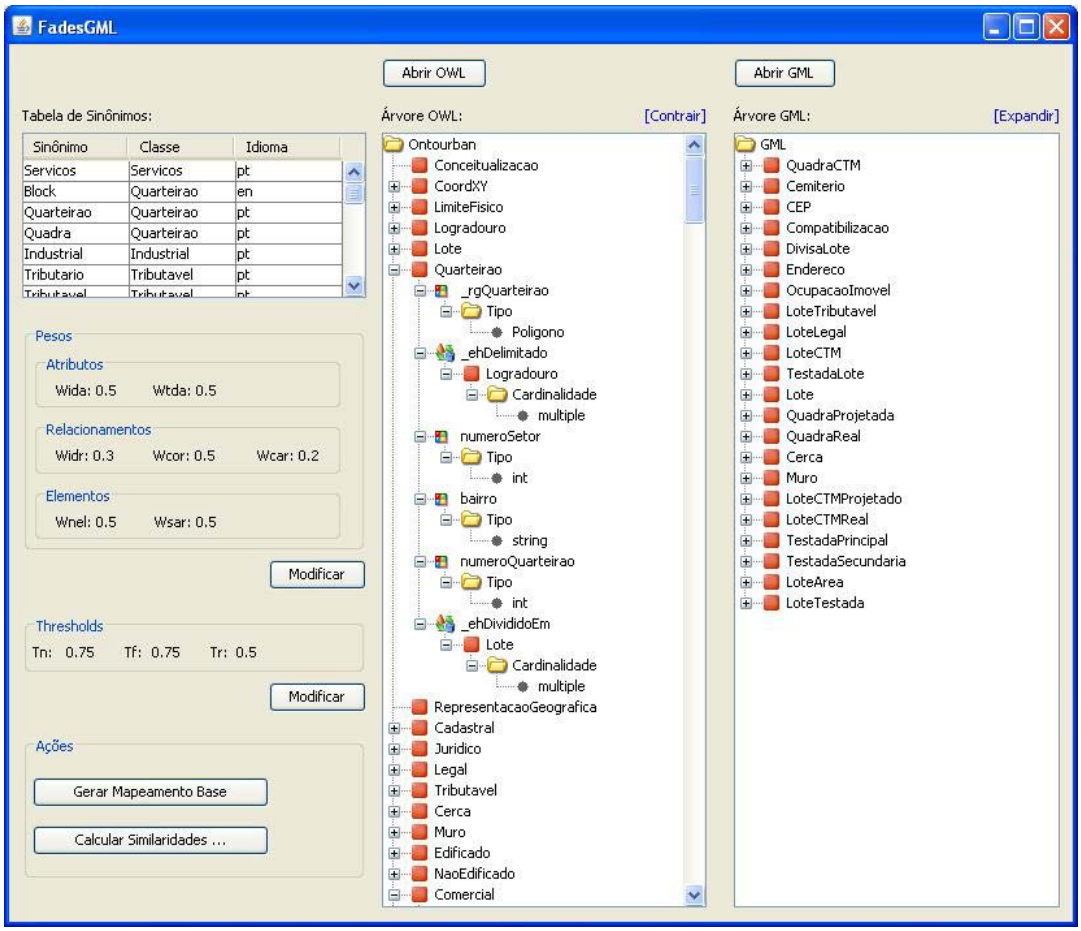


FIGURA 26: Interface principal da ferramenta

Conforme pode ser visto na Figura 26, há dois grandes quadros, um que apresenta a árvore canônica OWL e outro que apresenta a árvore canônica GML. Nestes quadros é possível expandir a estrutura das árvores, a fim de possibilitar a visualização dos componentes de cada elemento.

À esquerda, no alto, aparece a tabela de sinônimos com suas três colunas: *Sinônimo*, *Classe* e *Idioma*. Logo abaixo, estão os parâmetros do sistema (pesos e *thresholds*) e os botões que permitem a alteração de seus valores:

- *Wida* – peso para a similaridade dos nomes de atributos na métrica *simName()*;
- *Wtda* – peso para a similaridade dos tipos de dados de atributos na métrica *simName*(). A soma de *Wida* e *Wtda* deve ser igual a 1;
- *Widr* – peso para a similaridade dos nomes de relacionamentos na métrica *simRel()*;
- *Wcor* – peso para a similaridade dos conceitos referenciados em um relacionamento na métrica *simRel()*;
- *Wcar* – peso para a similaridade das cardinalidades em relacionamentos na métrica *simRel()*. A soma de *Widr*, *Wcor* e *Wcar* deve ser igual a 1;
- *Wnel* – peso da similaridade dos identificadores de um par de elementos GML/OWL na métrica *tupleSim()*;
- *Wsar* – peso da similaridade das propriedades (atributos e relacionamentos) de um par de elementos GML/OWL na métrica *tupleSim()*. A soma de *Wnel* e *Wsar* deve ser igual a 1;
- *Tn* – define o limite a partir do qual a equivalência entre nomes é considerada correta;
- *Tf* – define o limite a partir do qual a equivalência entre dois elementos GML/OWL é considerada correta;
- *Tr* – define o limite a partir do qual a equivalência de conceitos em um relacionamento é considerada provável.

Na base da interface, à esquerda, encontram-se dois botões de ação: *Gerar Mapeamento Base* e *Calcular Similaridades...*. O botão *Gerar Mapeamento Base*, como o nome indica, serve para a primeira execução do método, pela qual é feito o

mapeamento do esquema principal (GML_M) com a ontologia. Uma vez executada esta etapa, a ferramenta não permite mais fazer alterações nos parâmetros. Na prática, entende-se que um conjunto de parâmetros é definido uma única vez para o mapeamento do GML_M com a ontologia e sempre é reutilizado com os demais esquemas que se pretende obter o mapeamento. Se, em algum momento, o usuário optar por alterar os valores dos parâmetros, ele deve eliminar os mapeamentos armazenados no banco de dados e refazer todo o processo novamente. Na atual versão da ferramenta, tal eliminação é realizada toda vez que é selecionado o botão *Gerar Mapeamento Base*.

O segundo botão é usado para iniciar o processo que determina a equivalência dos esquemas importados (GML_I). Quando este botão é pressionado, a ferramenta solicita que o usuário informe os metadados do esquema GML_I, a fim de criar um registro identificador do esquema na tabela *EsquemaGMLi*. Na seqüência, após a execução do método, é gravado um registro na tabela *MapeamentoGMLi-GMLm* para cada equivalência definida como correta.

Ao final de cada execução da ferramenta, três telas são apresentadas em seqüência ao usuário:

a) *Resultado parcial* (Figura 27): apresenta as equivalências definidas automaticamente pela ferramenta;

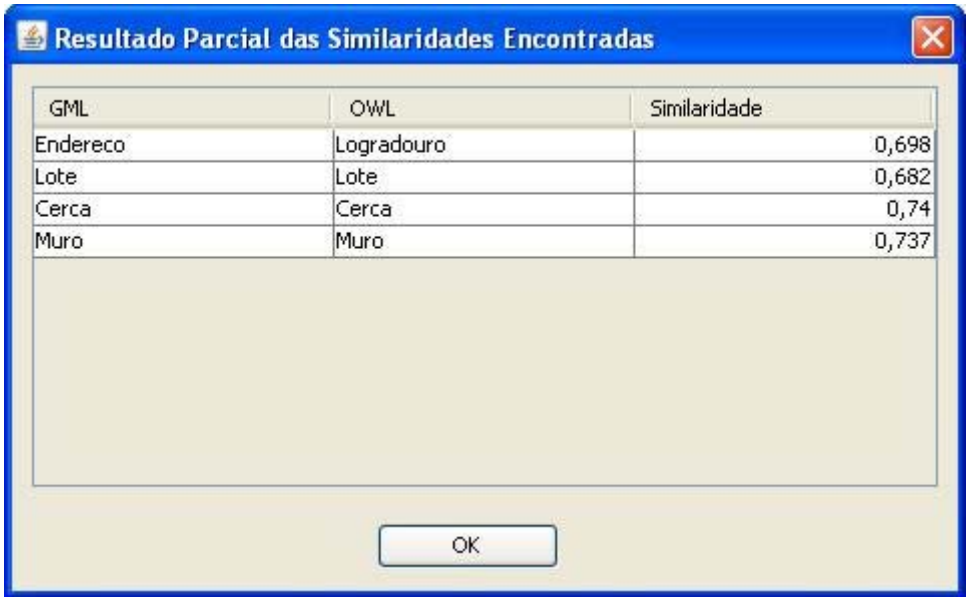


FIGURA 27: Tela dos resultados parciais

b) *Validação de equivalências* (Figura 28): apresenta os elementos GML para os quais foi calculada uma possível equivalência com mais de um conceito na ontologia. Nesta tela o usuário pode indicar quais pares de elementos GML/OWL são realmente equivalentes;

FIGURA 28: Tela para validação de equivalências

c) *Resultado final* (Figura 29): apresenta a lista de mapeamentos que são efetivamente armazenados no banco de dados.

FIGURA 29: Tela com os resultados finais

5.2 Dados de Entrada para o Estudo de Caso

De forma a validar o método proposto, foram criados três esquemas GML, incluindo o esquema GML descrito no capítulo 3, mais a ontologia com conceitos relacionados a lotes e quadras. Os esquemas GML correspondem a:

- *MUB-BH*: esquema GML representando parte do MUB-BH, criado a partir das classes encontradas nos pacotes *Quadras* (Figura 6) e *Lotes* (Apêndice III) definidos em Bertini (2003);
- *Lages*: esquema GML criado a partir dos esquemas conceitual (Apêndice V) e lógico fornecidos pela Prefeitura do Município de Lages;
- *CTM*: esquema criado a partir de exemplo de esquema conceitual encontrado em Borges, Davis Júnior e Laender (2005) (Apêndice VI).

A ontologia foi adaptada do modelo conceitual de ontologia de lote urbano proposto por Pinho e Goltz (2003), o qual tem por base a cidade de Belo Horizonte.

O esquema *MUB-BH* foi escolhido para ser o esquema principal (GML_M), uma vez que a origem do esquema conceitual proposto por Bertini (2003) também é a cidade de Belo Horizonte.

O Quadro 11 apresenta a quantidade de elementos, atributos e relacionamentos definidos na ontologia e em cada esquema GML. Para complementar, foi montado um dicionário contendo 63 sinônimos com variações dos nomes usados para as classes presentes na ontologia.

QUADRO 11: Quantidade de elementos e atributos da ontologia e dos esquemas GML

	Ontologia	***MUB-BH***	***Lages***	***CTM***
Conceitos/Elementos	25	22	39	16
Atributos	121	58	405	40
Relacionamentos	49	57	110	29

5.3 Exemplo de Aplicação das Métricas de Similaridade

Esta seção tem a finalidade de demonstrar a aplicação das métricas de similaridade utilizadas no método de determinação de equivalências semânticas e apresentadas na seção 4.4. Para tanto, é apresentado o cálculo da equivalência entre o elemento GML *QuadraCTM*, presente no esquema *MUB-BH*, e o conceito *Quarteirao*,

presente na ontologia. Os parâmetros configurados na ferramenta para esse exemplo são apresentados na Figura 30.

Na Fase 1 do processamento (ver seção 4.3), o método faz a procura por sinônimos equivalentes ao nome do elemento *QuadraCTM*. Como resultado, não é encontrado nenhum nome idêntico a *QuadraCTM*. Entretanto, são encontrados dois sinônimos equivalentes (*Quarteirao* e *Quadra*), associados à classe *Quarteirao*, ou seja, sinônimos cujo resultado de *simName()* (equação 5) é maior ou igual ao *threshold* de nomes (*Tn*). Como os dois sinônimos equivalentes encontrados referem-se à mesma classe, no restante do processamento utiliza-se o sinônimo que apresenta o maior valor de similaridade (*Quadra* = 0.95555556).

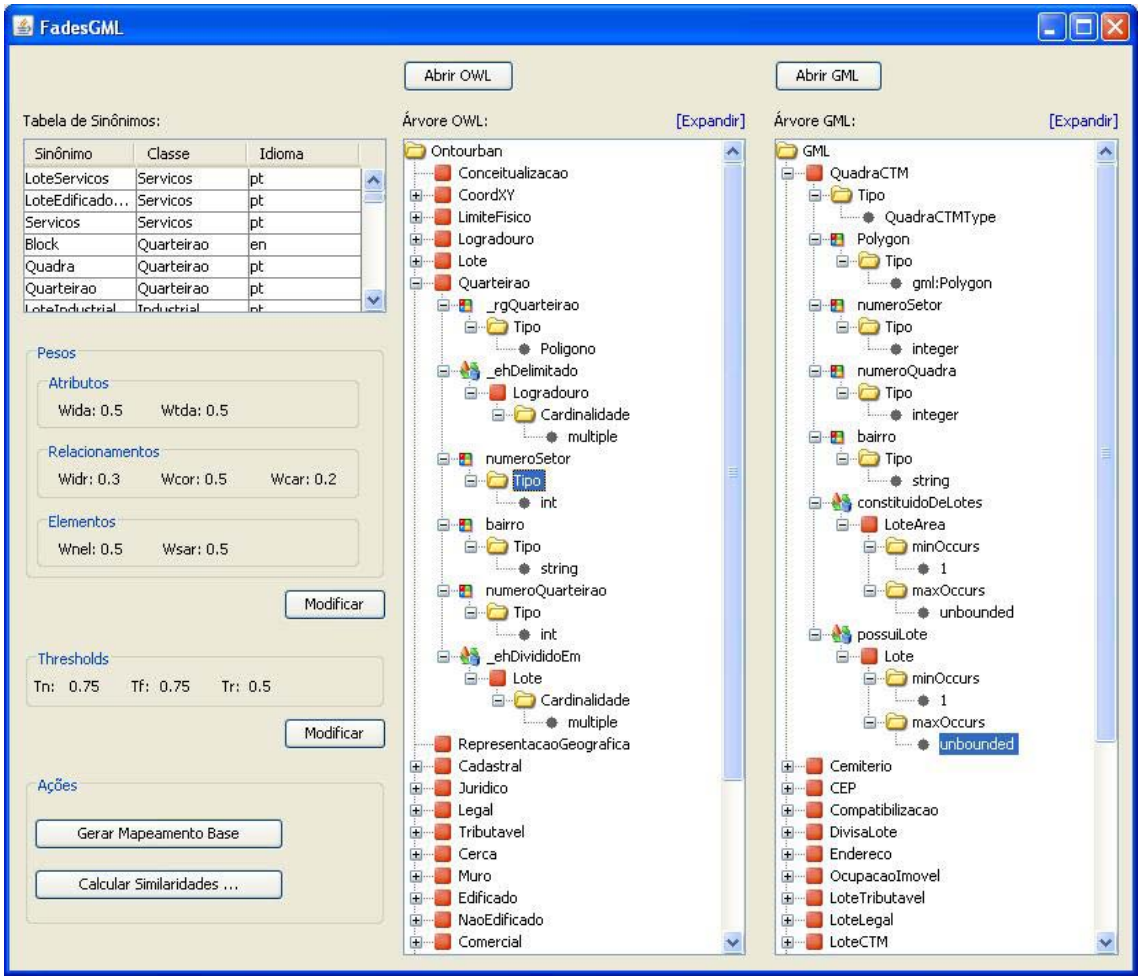


FIGURA 30: Interface da ferramenta com os dados para o exemplo de aplicação do método

Já na Fase 2, são calculadas inicialmente as similaridades dos atributos componentes dos dois elementos ($\sum simAttr()$), conforme mostra a Tabela 2. A similaridade entre dois atributos (equação 2) é igual à soma das similaridades do nome e

do tipo de dado dos atributos, levando em consideração também os respectivos pesos para nomes de atributos e tipos de dados (*Wida*, *Wtda*). Um par de atributos é considerado, para efeitos de cálculo, somente quando a similaridade de seus nomes for igual ou superior a *Tn*. Uma vez que um par de atributos é considerado equivalente, os mesmos não participam mais no cálculo dos demais atributos GML. Também é armazenado o maior número de atributos entre os dois elementos analisados (quatro atributos, neste caso).

TABELA 2: Similaridade dos atributos de *QuadraCTM* e *Quarteirao*

Atributos de *QuadraCTM*	**Atributos de *Quarteirao***	***simName() + simDataType()***
Polygon	*_rgQuarteirao*	0.0
numeroSetor	*numeroSetor*	(0.5 * 1.0) + (0.5 * 0.9) = 0.95
numeroQuadra	*numeroQuarteirao*	(0.5 * 0.94722223) + (0.5 * 0.9) = 0.92361111
Bairro	*bairro*	(0.5 * 1.0) + (0.5 * 1.0) = 1.0
∑simAttr()		**2.873611**

Ainda na Fase 2, é calculada a similaridade dos relacionamentos existentes nos dois elementos (***∑simRel()***), conforme mostra a Tabela 3.

TABELA 3: Similaridade dos relacionamentos de *QuadraCTM* e *Quarteirao*

Relacionamentos de		***simName() + simConc()+simCard()***
QuadraCTM	***Quarteirao***	
constituidoDeLotes	*_ehDelimitado*	(0.3 * 0.48746443) + (0.5 * 0.3533333) + (0.2 * 1.0) = 0.0
constituidoDeLotes	*_ehDivididoEm*	(0.3 * 0.5109481) + (0.5 * 0.5524012) + (0.2 * 1.0) = 0.62948503
possuiLote	*_ehDelimitado*	(0.3 * 0.48589745) + (0.5 * 0.3266666) + (0.2 * 1.0) = 0.50910253
possuiLote	*_ehDivididoEm*	(0.3 * 0.45128205) + (0.5 * 0.5496234) + (0.2 * 1.0) = 0.610196315
∑simRel()		**1.239681**

A similaridade entre dois relacionamentos (equação 3) é obtida pela soma das similaridades de seus nomes, dos conceitos relacionados e de suas cardinalidades, levando em consideração os respectivos pesos (*Widr*, *Wcor*, *Wcar*). Somente são considerados candidatos a relacionamentos equivalentes aqueles pares cuja similaridade dos conceitos relacionados for igual ou superior a *Tr* (conforme mostram as linhas em destaque nas Tabelas 3 e 4).

A similaridade dos conceitos relacionados (equação 4) é obtida pela soma da similaridade dos nomes dos conceitos e do resultado da soma das similaridades dos atributos dividida pelo maior número de atributos, levando em consideração os pesos *Wnel* e *Wsar* (Tabela 4).

TABELA 4: Cálculo da similaridade dos conceitos relacionados

Relacionamentos			***Similaridades dos conceitos (simConc())***		
Nome	***Conceito***	**Cardinalidade**	***simName()***	***∑simAttr()***	***Total***
constituidoDeLotes	*LoteArea*	*unbounded*	0.70666665	(0.0 / 5)	0.3533332
_ehDelimitado	*Logradouro*	*multiple*			
constituidoDeLotes	*LoteArea*	*unbounded*	0.9	(1.8432217 / 9)	0.5524012
_ehDivididoEm	*Lote*	*multiple*			
possuiLote	*Lote*	*unbounded*	0.6533333	(0.0 / 3)	0.3266666
_ehDelimitado	*Logradouro*	*multiple*			
possuiLote	*Lote*	*unbounded*	1.0	(0.8932217 / 9)	0.5496234
_ehDivididoEm	*Lote*	*multiple*			

Considera-se, também, que mais de um relacionamento do elemento GML pode ser equivalente a um mesmo relacionamento do conceito na ontologia. Por outro lado, quando um relacionamento em um elemento GML encontra mais de um candidato a equivalente no conceito da ontologia, então se escolhe como equivalente o par que tiver maior similaridade de nomes.

Finalmente, na Fase 3, aplica-se a *tupleSim()* (equação 1) para calcular a similaridade final dos elementos sendo analisados. Os dados disponíveis para este cálculo são:

- *Wnel* = 0.5
- *simName()* = 0.95555556
- *Wsar* = 0.5
- Similaridades das propriedades = $\sum simAttr()$ + $\sum simRel()$ = 2.873611 + 1.239681 = 4.113292
- Maior número de propriedades = 6

Assim, o resultado obtido com a *tupleSim()* é 0.820. Como este valor é superior ao *threshold* final (*Tf*), então se considera que o elemento GML *QuadraCTM* é equivalente ao conceito *Quarteirao* da ontologia. Desse modo, este mapeamento pode ser gravado no banco de dados.

5.4 Validação do Método

Para validar os resultados do método foram utilizadas duas medidas probabilísticas, amplamente usadas para validar algoritmos computacionais: *recall* e *precision* (SALTON, 1975).

O recall representa a proporção de material relevante recuperado e a *precision* representa a proporção de material recuperado que é relevante. No contexto deste trabalho, entende-se como material relevante recuperado cada par (mapeamento) formado por um elemento do esquema GML importado e uma classe da ontologia que seja corretamente (semanticamente) equivalente. Assim, o *recall* corresponde ao percentual de mapeamentos corretos em relação ao conjunto de mapeamentos esperados e a *precision* representa o percentual de mapeamentos corretos em relação ao total de mapeamentos obtidos. Com isso, pode-se dizer que um *sistema perfeito* é aquele cujo resultado apresenta um alto valor para *recall* e *precision*.

O *recall* (*R*) é definido pela equação 8:

$$R = \frac{\text{número de itens recuperados e relevantes}}{\text{total de itens relevantes na coleção}} \qquad \textbf{Eq. 8}$$

A *precision* (*P*) é definida pela equação 9:

$$P = \frac{\text{número de itens recuperados e relevantes}}{\text{total de itens recuperados na coleção}} \qquad \textbf{Eq. 9}$$

Para cada esquema GML utilizado, o Quadro 12 apresenta o número de elementos que possuem uma classe equivalente na ontologia e o número de classes na ontologia que possuem um elemento GML equivalente. Para definir estes valores, foram comparados os esquemas conceituais que deram origem aos esquemas GML e o esquema conceitual da ontologia. Foram considerados todos os elementos GML que pudessem ser equivalentes a uma classe na ontologia e todas as classes na ontologia que pudessem ser equivalentes a algum elemento GML.

QUADRO 12: Número de elementos GML e classes OWL com possíveis equivalências

	MUB-BH	***Lages***	***CTM***
Elementos GML com equivalência	16	09	06
Classes OWL equivalentes	43	37	42

O número de equivalências esperadas, ou seja, o número de itens relevantes na coleção, corresponde aos valores da segunda linha (*Classes OWL equivalentes*). A diferença entre os números apresentados para as duas linhas do Quadro 12 dá-se em função das especializações existentes na ontologia e nos esquemas GML, ou seja, essa diferença é conseqüência das cardinalidades 1:*n* e *n*:1. Por exemplo, o elemento GML

Lote está presente nos esquemas *MUB-BH* e *Lages*, enquanto que no esquema *CTM* ele é denominado *LoteCTM*. Estes elementos são equivalentes à classe *Lote* na ontologia, que possui outras nove subclasses (*NaoEdificado*, *Edificado*, *Serviços*, *Misto*, *Industrial*, *Comercial*, *Residencial*, *Uni-Familiar*, *Multi-Familiar*). Assim, inicialmente, para cada elemento GML indicado acima, espera-se encontrar um total de dez equivalências. Os elementos *Lote* nos esquemas *MUB-BH* e *Lages* possuem duas especializações cada, o que eleva o número de equivalências esperadas para 30 em cada esquema. O elemento *LoteCTM*, por sua vez, possui três especializações, gerando uma expectativa de 40 equivalências.

5.5 Resultados Obtidos

O método foi aplicado aos três esquemas GML indicados anteriormente, sendo que o esquema da cidade de Belo Horizonte (*MUB-BH*) corresponde ao esquema GML_M, enquanto os demais correspondem a esquemas GML_I.

Três cenários de execução distintos foram criados com o objetivo de analisar diferentes combinações de pesos e *thresholds*:

- *Cenário 1*: procurou-se anular a importância dos pesos em cada métrica e fixar os *thresholds* aleatoriamente em 0.5;
- *Cenário 2*: procurou-se definir pesos de maneira que se obtivesse o melhor mapeamento entre o esquema GML_M e a ontologia, em termos de maior número de mapeamentos corretos (*recall*);
- *Cenário 3*: os pesos de cada característica (nomes, tipos de dados, cardinalidades etc.) foram definidos de forma que melhor refletissem a sua importância na determinação semântica da equivalência.

Na mudança da execução de um cenário para outro, o banco de dados foi reinicializado, a fim de considerar apenas os mapeamentos obtidos com os novos valores dos parâmetros.

O Quadro 13 apresenta as configurações de pesos e *thresholds* usadas nos três cenários em que a ferramenta foi executada, enquanto que o Quadro 14 resume os resultados obtidos em cada um dos três cenários.

QUADRO 13: Pesos e *thresholds* configurados nos três cenários avaliados

	Tn	***Tf***	***Tr***	***Wida***	***Wtda***	***Widr***	***Wcor***	***Wcar***	***Wnel***	***Wsar***
Cenário 1	0.5	0.5	0.5	0.5	0.5	0.35	0.35	0.3	0.5	0.5
Cenário 2	0.75	0.5	0.5	0.7	0.3	0.3	0.5	0.2	0.7	0.3
Cenário 3	0.75	0.5	0.5	0.6	0.4	0.4	0.5	0.1	0.6	0.4

QUADRO 14: Resultados obtidos nos três cenários avaliados

Esquema	**Cenário**	**Automático (a)**	**Automático corretos (b)**	**Validado (c)**	**Não-Validado (d)**	**Total encontrado (a+c+d)**	**Total mapeado (a+c)**
MUB-BH	1º.	02	01	36	03	41	38
	2º.	03	03	35	52	90	38
	3º.	03	03	33	52	88	36
Lages	1º.	04	01	24	30	58	28
	2º.	02	01	25	31	58	27
	3º.	02	01	25	31	58	27
CTM	1º.	03	02	20	0	23	23
	2º.	02	02	40	0	42	42
	3º.	02	02	40	0	42	42

O Quadro 14 possui as seguintes informações: o total de equivalências definidas automaticamente; o total de resultados automáticos corretos; o total de resultados validados pelo usuário; o total não validado; o total de equivalências encontradas pelo método e o total de equivalências mapeadas (automáticas + validadas).

Ressalta-se que todas as equivalências automáticas são mapeadas atualmente pela ferramenta, porém alguns resultados encontrados apresentaram-se incorretos, o que levanta uma discussão sobre a possibilidade de o especialista também validar estas equivalências.

A Figura 31 apresenta o gráfico *recall* x *precision* para os resultados obtidos nos três cenários.

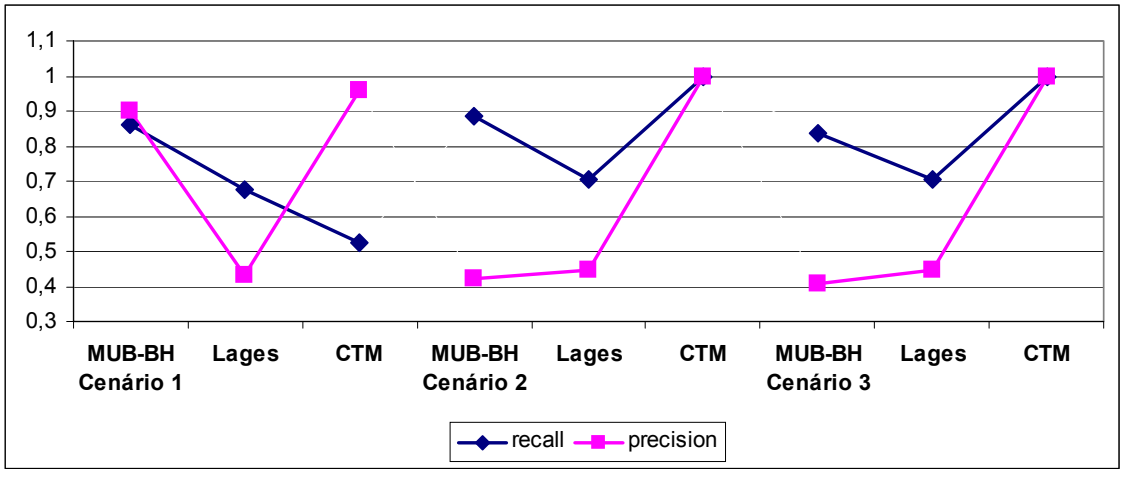


FIGURA 31: Gráfico *recall* x *precision* dos resultados obtidos

Em uma primeira análise, levando em consideração todos os testes realizados, para os esquemas *Lages* e *CTM*, os cenários 2 e 3 não apresentaram diferenças em termos de configuração de pesos e *thresholds*. Para o esquema *CTM* a ferramenta retornou todas as equivalências esperadas, enquanto que para o esquema *Lages*, melhores resultados só podem ser obtidos com mudanças no esquema ou na ontologia.

A partir de uma análise na estrutura dos três esquemas e da ontologia, verifica-se uma diferença muito grande em termos de quantidade e nomes de atributos e relacionamentos de cada elemento. Por exemplo, o conceito *Lote*, na ontologia, possui 9 atributos e 4 relacionamentos; no esquema *MUB-BH*, *Lote* possui 3 atributos e 5 relacionamentos; no esquema *Lages*, *Lote* possui 29 atributos e 8 relacionamentos; no esquema *CTM*, *LoteCTM* possui 2 atributos e 1 relacionamento. As diferenças de nomenclatura podem ser vistas na Figura 32.

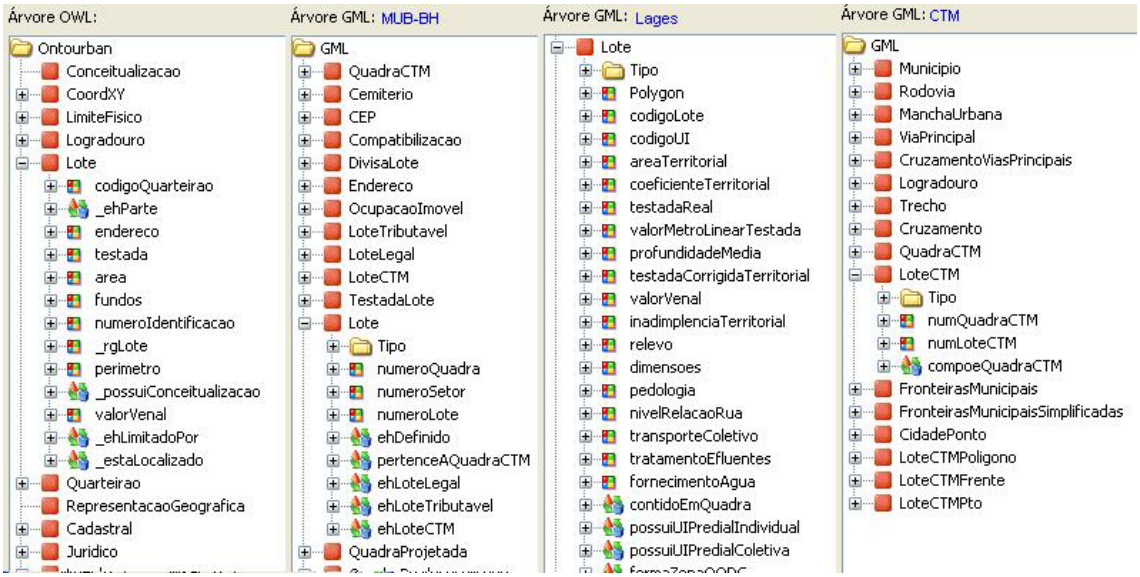


FIGURA 32: Exemplo das diferenças de quantidade e nome de atributos e relacionamentos

Essas diferenças têm grande influência na forma como os pesos usados nas métricas são definidos. Nos testes realizados, os melhores resultados foram obtidos quando se definiu um peso maior para a similaridade dos nomes (*Wida*, *Widr* e *Wnel*), em detrimento ao restante da estrutura (*Wtda*, *Wcor*, *Wcar* e *Wsar*). Como forma de compensar o valor atribuído ao peso dos nomes, o *threshold* (*Tn*) também foi mantido em um valor alto (0.75).

No que diz respeito aos relacionamentos, procurou-se estabelecer a similaridade semântica principalmente em função dos conceitos relacionados e, em

menor escala, em relação aos nomes dos relacionamentos. Percebe-se, na base de testes em questão, que a cardinalidade tem pouca influência, podendo até mesmo ter seu peso definido em zero para a base de testes utilizada.

A partir do conjunto de resultados obtidos, observa-se que aqueles mapeamentos em que há uma proximidade muito grande entre nomenclaturas e estrutura (como, por exemplo, a classe *Quadra*), a ferramenta obteve um índice de equivalência alto, ou seja, superior a 0.7.

Outra observação importante a respeito da base de testes usada é que a ferramenta retornou um grande número de pares de elementos para validação. Boa parte desses pares foi validada, uma vez que se referiam a situações em que um elemento GML era equivalente a uma classe da ontologia e, também, às suas diversas especializações. As equivalências obtidas na maioria desses casos tiveram uma variação entre 0.55 e 0.70.

Do conjunto de equivalências apresentadas e não validadas cabe uma observação a respeito dos esquemas *MUB-BH* e *Lages*. No esquema *MUB-BH* existe um conjunto de elementos (*Compatibilizacao*, *LoteTributavel*, *LoteLegal*, *LoteCTM*, *LoteCTMProjetado* e *LoteCTMReal*) usados para representar as várias definições de *Lote*, conforme exposto no capítulo 3. Porém, a ferramenta encontrou similaridades destes elementos com algumas das classes que representam especializações de *Lote* na ontologia. Essas equivalências foram fortemente influenciadas pelo peso atribuído à similaridade do nome de elementos (*Wnel*) na métrica *tupleSim()*. No esquema *Lages*, por sua vez, existe um conjunto de elementos especializados de *Loteamento* (*LoteamentoReal* e *LoteamentoProjetado*) que, pela similaridade dos nomes, a ferramenta atribuiu equivalência com o conceito *Lote*. Todos esses casos dependem de uma análise do especialista que deve definir se podem ou não ser considerados equivalentes. A faixa de equivalências obtida nestes casos variou entre 0.5 e 0.62.

Concluindo a análise sobre os dois últimos parágrafos, o método mostrou-se bastante eficaz em relação ao mapeamento de elementos que não deixam dúvidas sobre sua equivalência, retornando valores acima de 0.7, com um índice de erro muito pequeno. Acredita-se que os resultados comentados nos dois últimos parágrafos podem

ser melhorados com duas ações sobre a configuração inicial da ferramenta, realizadas de forma conjunta:

a) revisão da ontologia, principalmente em relação à quantidade de atributos e relacionamentos nas classes, bem como em relação à qualidade das nomenclaturas. Em alguns casos, a inclusão de novos sinônimos pode melhorar a quantidade de equivalências corretas obtidas;
b) definição de pesos que valorizem mais a estrutura, principalmente *Wtda* e *Wsar*.

Com isso, espera-se reduzir as grandes variações nos valores finais obtidos devido, principalmente, às diferenças na quantidade de atributos entre um elemento GML e a respectiva classe na ontologia.

Os resultados para *recall* e *precision* obtidos para cada esquema GML nos três cenários em que a ferramenta foi executada são apresentados na Tabela 5.

TABELA 5: Resultados de *recall* e *precision*

		MUB-BH	***Lages***	***CTM***
Cenário 1	***recall***	0.8605	0.6757	0.5238
	precision	0.9024	0.4310	0.9565
Cenário 2	***recall***	0.8837	0.7027	1.0000
	precision	0.4222	0.4483	1,0000
Cenário 3	***recall***	0.8372	0.7027	1,0000
	precision	0.4091	0.4483	1.0000

Como pode ser visto na Tabela 5, a primeira execução da ferramenta, cujos parâmetros visam dar importância igual para os componentes dos elementos, obteve melhores resultados de *recall* e *precision* para o esquema *MUB-BH*, uma vez que os elementos deste esquema são muito parecidos com as classes da ontologia e, neste caso, as nomenclaturas tiveram uma grande influência no resultado. O esquema *CTM* teve um grau de *precision* também elevado. Tal resultado era esperado, uma vez que a origem das nomenclaturas do esquema (e seus autores) é a mesma do esquema *MUB-BH*.

Na segunda execução, cujo objetivo foi encontrar o maior número de mapeamentos corretos entre a ontologia e o esquema *MUB-BH*, este esquema também obteve um bom resultado para *recall*, porém, a precisão foi menor, já que apresentou um elevado número de mapeamentos para serem validados. O esquema *CTM*, por sua

vez, apresentou um bom resultado, porém ressalta-se a observação quanto à origem do mesmo, feita no parágrafo anterior.

A terceira execução, que explorou os pesos de forma a valorizar melhor a semântica dos componentes, obteve resultados satisfatórios para os três esquemas. A ferramenta apresentou um número menor de mapeamentos para o esquema *MUB-BH* (Quadro 14), porém a qualidade dos mesmos foi melhor. Os esquemas *Lages* e *CTM* não apresentaram diferenças em relação à segunda execução, o que demonstra que a ferramenta está configurada em seus valores ótimos para estes esquemas.

Para os esquemas utilizados, outras mudanças nos pesos de atributos e relacionamentos (*Wida*, *Wtda*, *Widr*, *Wcor* e *Wcar*) além das definidas no Quadro 13, têm pouca influência no resultado final, apenas melhorando um pouco os valores de similaridade calculados. Por outro lado, mudanças nos pesos dos elementos (*Wnel* e *Wsar*), assim como nos *thresholds* (*Tn*, *Tf* e *Tr*) têm grande influência nos resultados finais, o que denota a forte importância das nomenclaturas nos esquemas analisados. Como já comentado anteriormente, a atualização da ontologia pode refletir em melhores resultados, principalmente em relação a atributos e relacionamentos.

Em sistemas de integração de dados, considera-se que o *recall* é o parâmetro mais importante, uma vez que ele determina a quantidade de resultados corretos que foram obtidos. Neste sentido, a ferramenta demonstrou que atende a esta necessidade.

Com relação à *precision*, ela indica quanto o especialista é exigido para confirmar os resultados obtidos, ou seja, quanto maior a *precision*, menor o trabalho do especialista. Para os esquemas *MUB-BH* e *Lages*, a *precision* apresentou um valor baixo por causa do grande número de resultados apresentados para validação do especialista, porém não eram equivalências corretas, conforme discutido anteriormente. Para melhorar este resultado, um ponto a ser aperfeiçoado é a forma de tratar as especializações, buscando valorizá-las mais em relação aos demais elementos, e aumentar a quantidade de mapeamentos automáticos definidos pela ferramenta.

5.6 Conclusão

Este capítulo apresentou o estudo de caso realizado para a validação do método proposto no capítulo 4, o qual compreende a descrição da ferramenta implementada

para este propósito, a origem dos esquemas que compõem a base de testes e a apresentação dos resultados obtidos com a ferramenta.

Quanto à ferramenta, ela implementa o algoritmo apresentado na seção 4.3 e as métricas apresentadas na seção 4.4. A interface criada permite a visualização dos dados de entrada, a navegação nas árvores canônicas dos esquemas GML e da ontologia, a comparação visual de um elemento GML contra uma classe na ontologia pela expansão das árvores e, por fim, a verificação dos mapeamentos definidos automaticamente e a validação dos mapeamentos 1:*n* e *n*:1 por um especialista.

Com relação aos dados de entrada, os três esquemas GML usados apresentam grandes diferenças em relação a nomenclaturas e estrutura. Apesar da origem dos autores dos esquemas *MUB-BH* e *CTM* ser a mesma, estes apresentam grandes semelhanças apenas na nomenclatura utilizada, sendo que a estrutura dos esquemas e a composição dos elementos (atributos e relacionamentos) é bem diferente.

Quanto à validação do método, foram criados três cenários distintos, os quais foram aplicados aos três esquemas disponíveis. Os resultados obtidos foram bastante satisfatórios, não apontando problemas no algoritmo ou nas métricas propostas. Por outro lado, percebe-se que algumas decisões são importantes no momento de configurar os pesos e *thresholds* na ferramenta. A principal decisão é definir se a intenção é obter um melhor mapeamento entre a ontologia e o esquema GML_M ou definir os parâmetros de forma que melhor tratem a semântica dos dados, sem considerar um esquema em particular. Neste sentido, percebe-se que grandes diferenças entre a estrutura de um elemento GML e de uma classe na ontologia têm forte influência no resultado final, ou seja, quando há uma grande diferença em termos de número de atributos e relacionamentos.

Como explicado anteriormente, não foi possível obter dados reais para testar a aplicação, sendo que os esquemas GML e a ontologia foram derivados de esquemas conceituais de banco de dados encontrados nas referências bibliográficas ou, no caso do esquema conceitual de Lages, disponibilizado por uma única organização. Esse fato pode ser considerado uma limitação do trabalho, mas, por outro lado, abre espaço para que novas oportunidades sejam exploradas, em parceria com organizações detentoras desses dados.

6 CONCLUSÃO E TRABALHOS FUTUROS

Os avanços que vêm acontecendo nas últimas décadas, na área de Informática e Computação, possibilitaram a popularização de várias tecnologias, entre elas, os Sistemas de Informações Geográficas (SIGs). O crescimento do uso de SIGs tem gerado uma série de outras necessidades para a comunidade de usuários e desenvolvedores deste tipo de aplicação. Como o custo de produção de dados georreferenciados é muito elevado, o compartilhamento e a troca de dados entre organizações aparece como uma alternativa interessante, principalmente pelas facilidades do uso intenso de redes de computadores. Entretanto, devido à complexidade da representação dos dados geográficos e a forma como cada aplicação armazena seus dados, a troca de informações entre SIGs apresenta incompatibilidades nos níveis sintático e semântico. Esses fatos têm levado instituições de pesquisa (UCGIS, 2002) a apontar a interoperabilidade semântica como uma prioridade de pesquisa atual.

Esta dissertação insere-se no contexto da interoperabilidade em SIGs, tratando uma parte específica do problema, a determinação de equivalências semânticas entre esquemas GML (*Geography Markup Language*). Para tanto, faz uma revisão de alguns trabalhos relacionados, com o objetivo de identificar as técnicas utilizadas para a solução do problema da interoperabilidade e as tecnologias que se apresentam como tendências para este fim.

Com relação às técnicas, percebe-se o interesse no desenvolvimento de aplicações que façam a mediação entre os diversos formatos de armazenamento encontrados em SIGs distintos, produzindo o mapeamento entre os dados ou a tradução de consultas através do uso de uma base de conhecimento comum, principalmente ontologias.

Em se tratando de tecnologias, destaca-se que o uso de padrões abertos se mostra como principal tendência. Entre esses padrões está a GML, que vem sendo desenvolvida por um consórcio internacional (*Open Geospatial Consortium*), formado por representantes da indústria de aplicações para SIG, usuários e centros de pesquisa

(OGC, 2007). A GML é um padrão para o transporte e armazenamento de dados geográficos que segue os mesmos princípios da XML (*eXtensible Markup Language*), amplamente utilizada na atualidade para a interoperabilidade em sistemas de aplicação. Outra tecnologia de destaque é a OWL (*Web Ontology Language*), que é a especificação mais recente recomendada pelo W3C (*World Wide Web Consortium*) para a definição de ontologias.

A partir da identificação das tendências, propõe-se um método para determinar as equivalências semânticas entre esquemas GML apoiado por uma ontologia desenvolvida em OWL. O desenvolvimento do método teve por base o domínio do cadastro urbano, uma vez que este domínio é pouco explorado em trabalhos relacionados e apresenta grande potencial para aplicações práticas, principalmente aquelas que envolvem dados para planejamento urbano.

O método divide-se em três módulos principais: a) criação de uma representação canônica para a ontologia e os esquemas GML; b) aplicação do algoritmo para determinar as equivalências semânticas entre os elementos dos esquemas GML; c) mapeamento das equivalências encontradas com os elementos de um esquema GML principal.

Como formato de representação canônica dos elementos dos esquemas GML e dos conceitos presentes na ontologia, foi escolhida uma estrutura em árvore, forma adequada para representar e manipular dados XML, considerando que tanto a GML quanto a OWL são linguagens baseadas em XML. Esta estrutura de árvore foi definida com três níveis: a) *raiz*, que identifica o tipo de árvore (OWL ou GML); b) *elemento complexo*, que representa um elemento GML ou um conceito OWL; c) *componentes dos elementos complexos*, que representam os atributos e relacionamentos existentes nos elementos GML e nos conceitos OWL. Os componentes dos elementos complexos têm papel crucial na determinação da equivalência semântica.

O algoritmo de determinação de equivalências semânticas divide o problema em pequenas partes, para as quais aplica um conjunto de métricas de similaridades. Estas métricas seguem a classificação proposta por Dorneles *et al.* (2004), compreendendo: Métricas para Valores Complexos (MCV) e Métricas para Valores Atômicos (MAV). Uma das métricas MCV propostas por Dorneles *et al*. (2004), a

tupleSim(), foi adaptada para uso no contexto deste trabalho. Outras três métricas MCV foram criadas especificamente para o algoritmo proposto, além de três métricas MAV, que são dependentes do domínio da aplicação. Complementam o algoritmo, um conjunto de pesos e *thresholds*, utilizados para definir a importância de cada componente analisado no contexto semântico e delimitar o ponto a partir do qual os resultados podem ser considerados corretos.

Por fim, o mapeamento das equivalências é feito em um banco de dados relacional, primeiramente entre a ontologia e um esquema principal (GML_M) e, posteriormente, entre esquemas importados (GML_I) e o esquema GML_M. Alguns mapeamentos são definidos automaticamente pelo método proposto, enquanto que outros mapeamentos envolvem a participação do usuário especialista para confirmar se estão corretos.

Para validar a proposta, além de uma ontologia cobrindo os conceitos de *Lote* e *Quadra*, foram criados três esquemas GML a partir de esquemas conceituais de banco de dados geográficos, sendo dois obtidos nas referências bibliográficas e um cedido pela Prefeitura do Município de Lages – SC. Para os testes, foi implementada uma ferramenta através da qual foram realizadas diversas execuções do método. Ao final, foram relatados os resultados das execuções para cada esquema GML em três cenários diferentes: a) pesos definidos de forma a anular sua influência no resultado final; b) pesos definidos de forma a obter a maior quantidade de mapeamentos para o esquema GML_M; c) pesos definidos de forma a melhor representar a importância semântica de cada componente dos esquemas.

Os resultados obtidos com a ferramenta mostraram-se satisfatórios. Apesar de obter poucos mapeamentos automáticos, a ferramenta apresentou um grande número de mapeamentos para validação do usuário especialista. A maioria destas validações refere-se a mapeamentos de um elemento GML para várias classes na ontologia, os quais ocorrem quando uma classe na ontologia possui especializações. Poucos resultados realmente incorretos foram encontrados.

Algumas observações podem ser feitas a partir da execução dos diversos cenários na ferramenta. Primeiramente, a configuração dos parâmetros (pesos e *thresholds*) tem importância significativa no resultado final do método, porém, não

existe um conjunto de valores que possa ser indicado como correto. A melhor combinação vai depender do conjunto de dados de entrada (ontologia e esquemas) e do objetivo do usuário: obter um melhor mapeamento do esquema GML_M com a ontologia ou obter um mapeamento mais preciso, dando mais ênfase aos componentes semânticos.

Apesar de se concluir que os resultados obtidos superaram as expectativas, cabe salientar que o trabalho apresentado apresenta duas limitações. A primeira é que não foram encontrados dados reais para usar com o método, ou seja, tanto a ontologia quanto os esquemas GML foram criados especificamente para o trabalho. Essa limitação pode ter influência na validação do método, uma vez que uma ontologia ou esquema GML real podem mostrar detalhes que não foram percebidos até o momento como, por exemplo, padrões na definição de nomenclaturas e a quantidade de propriedades (atributos e relacionamentos) existentes em cada elemento. A segunda limitação vem da necessidade de reduzir o campo de atuação a um pequeno conjunto de conceitos do domínio cadastro urbano, ou seja, o método foi baseado nos conceitos *Lote* e *Quadra*. Entretanto, estes conceitos formam a principal camada de um Mapa Urbano Básico (MUB), o que justifica seu tratamento preferencial.

Assim, pode-se concluir que os objetivos propostos foram alcançados plenamente. Foi especificado um método para determinação de equivalências semânticas entre esquemas GML no domínio do cadastro urbano. Os tipos de equivalências tratadas estão descritos no texto e são apresentadas as métricas de similaridade desenvolvidas para este fim. O método proposto foi validado através de um estudo de caso, para o qual foi desenvolvida uma ontologia de domínio e três esquemas GML com características heterogêneas.

Como contribuições deste trabalho, pode-se destacar o uso em conjunto de uma ontologia OWL e de esquemas GML, os quais representam o estado-da-arte em termos de tecnologia para criação de bases de conhecimento e transporte de dados geográficos, respectivamente. Diferente de outros trabalhos, aqui propõe-se identificar as similaridades dos elementos de esquemas GML contra a ontologia, em vez de fazer diretamente entre dois esquemas GML. A vantagem dessa abordagem é que a ontologia representa melhor o conhecimento do domínio, podendo evoluir após sucessivas aplicações do método com diferentes esquemas GML. Além disso, pode-se ter

ontologias especializadas e mais adequadas para sub-áreas de aplicação ou sub-conjuntos de esquemas GML.

Soma-se a isso a preocupação em desenvolver soluções de SIG para o domínio do cadastro urbano, de maneira a fomentar o maior uso de dados geográficos para planejamento urbano. Do ponto de vista tecnológico, destaca-se a preocupação com o detalhamento do método proposto e das métricas de similaridade usadas, característica muitas vezes não encontrada em trabalhos relacionados. Acrescenta-se, ainda, a implementação de uma ferramenta para demonstração do método. Do ponto de vista acadêmico, o trabalho abre uma nova linha de pesquisa no Grupo de Banco de Dados, da Universidade Federal de Santa Catarina – UFSC: interoperabilidade em SIGs, uma vez que é o primeiro trabalho desenvolvido nesta área. Além disso, ele serve como importante referencial teórico para outros trabalhos que venham a tratar do mesmo assunto.

Com relação a trabalhos futuros, percebe-se um grande número de oportunidades para a continuidade desta pesquisa, além de outros trabalhos complementares. Câmara (2000, 2006) defende que dados públicos devem ser disponibilizados de forma aberta e que devem ser desenvolvidas ferramentas de baixo custo para acessá-los. Neste sentido, um maior estudo sobre a GML deve ser desenvolvido, com a criação de novas referências bibliográficas e ferramentas de desenvolvimento. Atualmente, *softwares* usados normalmente para validação de esquemas XML permitem apenas indicar se um esquema GML é válido enquanto esquema XML, mas não necessariamente se é válido observando-se todas as regras para esquemas GML. O modelo OMT-G mostra-se bastante adequado para a modelagem de dados geográficos e é amplamente usado, a ponto de a CONCAR - Comissão Nacional de Cartografia e a ANA - Agência Nacional de Águas, terem adotado o modelo como padrão. Porém, ainda não há uma ferramenta adequada para a criação de esquemas geográficos baseados neste modelo. Portanto, torna-se importante desenvolver um *software* para este fim e, melhor ainda, se tiver a possibilidade de traduzir esquemas OMT-G para esquemas GML.

Outro trabalho necessário é a definição de uma ontologia própria para o domínio do cadastro urbano, uma vez que a usada neste trabalho serve apenas como

modelo para testes. Essa ontologia pode ser criada de forma a manter a divisão dos conceitos conforme as camadas existentes em um MUB (CTM, Energia, Hidrografia, Transporte etc.). O projeto para desenvolvimento desta ontologia pode ser apresentado ao poder público, de forma a tornar-se uma referência nacional, que facilitaria posteriormente os processos de troca e compartilhamento de dados.

Especificamente sobre este trabalho, pesquisas futuras devem analisar a ampliação do dicionário de sinônimos, que atualmente só abrange sinônimos para as classes da ontologia. É interessante pesquisar a influência que sinônimos para os atributos podem exercer nos resultados do método. Como sugestão, um sinônimo para um atributo deve ser restrito a uma classe específica na ontologia e não tratado de forma igual para todas as classes. Outra sugestão considera a ampliação do escopo abordado, estendendo o método proposto para resolver as equivalências de novos conceitos e, também, novos domínios. A atualização da ontologia a partir das equivalências validadas pelo especialista pode ser tratada por meio de interfaces próprias. Também é possível especificar um ambiente para consulta a dados remotos, com base nas equivalências encontradas e, por fim, pode-se obter a integração de instâncias de dados geográficos. Com relação às equivalências *n*:*m*, atualmente elas existem principalmente devido às hierarquias de especialização que podem ocorrer na OWL e na GML. Isso acaba por não tornar muito direta (1:1) as equivalências. Uma solução a ser analisada é a possibilidade do especialista sempre escolher uma única correspondência em uma hierarquia de especialização, ou seja, aquele conceito OWL que melhor represente o elemento, e vice-versa. Assim, é possível obter mapeamentos mais precisos.

REFERÊNCIAS BIBLIOGRÁFICAS

AZEVEDO, V. H. M. *et al*. Interoperabilidade entre objetos geográficos heterogêneos. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA - GEOINFO, 8., 2006, Campos do Jordão. **Anais...** São José dos Campos: INPE, 2006.

BÉDARD, Y. *et al*. Adapting data models of the design on spatio-temporal databases. **Computers, Environment and Urban Systems**, v. 20, n. 1, p. 19-41, 1996.

BERTINI, G. da C. **Uma modelagem orientada a objeto para o mapa urbano básico de Belo Horizonte (MUB/BH)**. 58 f. 2003. Monografia (Especialização em Informática Pública) – Centro de Capacitação em Informática Pública, Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte.

BORGES, K. A. V. **Modelagem de Dados Geográficos: Uma Extensão do Modelo OMT para Aplicações Geográficas**. 139 f. 1997. Dissertação (Mestrado em Administração Pública) – Escola de Governo, Fundação João Pinheiro, Belo Horizonte.

BORGES, K. A. V.; DAVIS JUNIOR, C. A.; LAENDER, A. H. F. OMT-G: an object-oriented data model for geographic applications. **GeoInformatica**, v. 5, n. 3, p. 221-260, 2001.

BORGES, K. A. V.; DAVIS JÚNIOR, C. A.; LAENDER, A. H. F. Modelagem conceitual de dados geográficos. In: CASANOVA, M. A. *et al*. (org.) **Bancos de dados geográficos**. Curitiba: MundoGEO, 2005. p. 93-146.

BRAUNER, D. F.; CASANOVA, M. A.; LUCENA, C. J. P. Geo-Object catalogs to enable Geographic Databases Interoperability. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA - GEOINFO, 6., 2004, Campos do Jordão. **Anais...** São José dos Campos: INPE, 2004.

CÂMARA, G. **Análise Geográfica e Interoperabilidade**. Palestra. In: GEOBRASIL - CONGRESSO E FEIRA INTERNACIONAL DE GEOINFORMAÇÃO, 5., 2004, São Paulo, 2004.

CÂMARA, G. Entrevista: Gilberto Câmara. **InfoGeo,** São Paulo, n. 42, p. 24-27, jun. 2006.

CÂMARA, G. *et al*. Interoperability in Practice: Problems in Semantic Conversion from Current Technology to OpenGIS. In: INTERNATIONAL CONFERENCE ON INTEROPERABLE OPERATING SYSTEMS, 2., 1999. **Proceedings...** Zurich, 1999.

CÂMARA, G. *et al*. **Intercâmbio de Dados Geográficos no Brasil: um formato aberto**. Documento de Trabalho. Divisão de Processamento de Imagens - INPE. São José dos Campos: INPE, 2000. Disponível em: <http://www.dpi.inpe.br/geobr/>. Acessado em: 01 out. 2004.

CÂMARA, G. Representação computacional de dados geográficos. In: CASANOVA, M. A. *et al*. (org.) **Bancos de dados geográficos**. Curitiba: MundoGEO, 2005. p. 11-52.

CASANOVA, M. A. *et al.* Integração e interoperabilidade entre fontes de dados geográficos. In: CASANOVA, M. A. *et al*. (org.) **Bancos de dados geográficos**. Curitiba: MundoGEO, 2005. p. 317-352.

CHAPMAN, J. **Sam's Strings Metrics**. Apresenta uma biblioteca com implementações de métricas de similaridade. Disponível em: <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html>. Acessado em: 25 mai. 2006.

CRUZ, I. F.; RAJENDRAN, A. Semantic Data Integration in Hierarchical Domains. **IEEE Intelligent Systems**. v. 8, n. 2, p. 66-73, 2003.

DAVIS JUNIOR, C. A. *et al.* O Open Geospatial Consortium. In: CASANOVA, M. A. *et al*. (org.) **Bancos de dados geográficos**. Curitiba: MundoGEO, 2005. p. 379-395.

DOM. **Document Object Model**. Disponível em: <http://www.w3.org/DOM/>. Acessado em: 10 jan. 2007.

DORNELES, C. F. *et al*. Measuring Similarity between Collection of Values. In: ACM INTERNATIONAL WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT (WIDM), 6., 2004. **Proceedings...** Washington DC: ACM, 2004.

ELMASRI, R.; NAVATHE, S. **Fundamentals of database systems**. Pearson Education, 2004.

ESRI. **Shapefile Technical Description**. Environmental System Research Institute, Inc. Disponível em: <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>. Acessado em: 02 mar. 2005.

FERRARI, R. **Viagem ao SIG - Planejamento Estratégico, Viabilização, Implantação e Gerenciamento de Sistemas de Informação Geográfica**. Curitiba: Sagres, 1997.

FONSECA, F.; EGENHOFER, M. Ontology-Driven Geographic Information Systems. In: INTERNATIONAL SYMPOSIUM ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS – ACM GIS, 7., 1999, Kansas City. **Proceedings…** Kansas City: ACM, 1999. p. 14-19.

FONSECA, F.; EGENHOFER, M.; BORGES, K. Ontologias e Interoperabilidade Semântica entre SIGs. In: WORKSHOP BRASILEIRO EM GEOINFORMÁTICA, 2., 2000, São Paulo. **Anais...** São Paulo: SBC, 2000.

FORNARI, M. R.; IOCHPE, C. Mapping of Conceptual Object Oriented Models to Geography Markup Language (GML). In: IADIS INTERNATIONAL CONFERENCE WWW/INTERNET, Lisboa, 2002. **Proceedings...** Lisboa: IADIS, 2002. p. 444-451.

FROZZA, A. A.; MELLO, R. dos S. Determinando equivalências semânticas entre componentes de Esquemas GML. In: ESCOLA REGIONAL DE BANCO DE DADOS - ERBD, 2., 2006, Passo Fundo. **Anais...** Passo Fundo: SBC/UPF, 2006a. p. 63-68.

FROZZA, A. A.; MELLO, R. dos S. Um Método para Determinar a Equivalência Semântica entre Esquemas GML. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA - GEOINFO, 8., 2006, Campos do Jordão. **Anais...** São José dos Campos: INPE, 2006b. p. 283-294.

FROZZA, A. A.; MELLO, R. S. A Method for Defining Semantic Similarities between GML Schemas. In: Clodoveu Davis (Org.). **Advances in Geoinformatics**. Berlin: Springer-Verlag, 2007. (no prelo)

GAHEGAN, M. Accounting for the semantic differences between various Geographic Information Systems. In: INTERNATIONAL CONFERENCE AND WORKSHOP ON INTEROPERATING GEOGRAPHIC INFORMATION SYSTEMS, 1997, Santa Barbara (EUA). **Proceedings…** Santa Barbara: NCGIA, 1997.

GOMES, A. C. dos R. **A representação do lote CTM no geoprocessamento de Belo Horizonte**. 45 f. 2000. Monografia (Especialização em Geoprocessamento) – Departamento de Cartografia, Universidade Federal de Minas Gerais, Belo Horizonte.

HESS, G. N. **Unificação semântica de esquemas conceituais de bancos de dados geográficos**. 110 f. 2004. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre.

HESS; G. N.; IOCHPE, C. Utilizando a GML na identificação de candidatos a padrão de análise para BDG. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA – GEOINFO, 5., Campos do Jordão, 2003. **Anais...** São José dos Campos: INPE, 2003.

HESS, G. N.; IOCHPE, C. Ontology-Driven Resolution of Semantic Heterogeneities in GDB Conceptual Schemas. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA – GEOINFO, 6., Campos do Jordão, 2004. **Anais...** São José dos Campos: INPE, 2004.

HESS, G. N.; IOCHPE, C.; CASTANO, S. An Algorithm and Implementation for GeoOntologies Integration. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA – GEOINFO, 8., Campos do Jordão, 2006. **Anais...** São José dos Campos: INPE, 2006.

HOHL, P. **GIS Data Conversion: Strategies, Techniques and Management**. New York: Onword Press, 1998.

IBGE. **Censo Demográfico 2000**: População. Disponível em: <http://www.ibge.gov.br/>. Acessado em: 22 fev. 2004.

KASHYAP, V.; SHETH, A. P. Semantic and Schematic Similarities Between Database Objects: A Context-Based Approach. **The VLDB Journal**, v. 5, n. 4, p. 276-304, 1996.

KÖSTERS, G.; PAGEL, B.; SIX, H. GIS Application Development with GeoOOA. **International Journal of Geographic Information Science**, v. 11, n. 4, p. 307-335, 1997.

LAKE, R. Building the Geo-web at the local, regional, national and global levels. In: DATABASE 2002 CONFERENCE, Out., 2002, Tokyo. **Proceedings…**

LIMA, P. **Intercâmbio de dados espaciais: modelos, formatos e conversores**. 2002. 79 f. Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais – INPE, São José dos Campos.

LISBOA FILHO, J. **Modelos de Dados Conceituais para Sistemas de Informação Geográfica**. 1997. 67 f. Exame de Qualificação (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre.

MOROCHO, V.; PÉREZ-VIDAL, L.; SALTOR, F. Semantic Integration on Spatial Databases: SIT-SD prototype. In: JORNADAS DE INGENIERÍA DEL SOFTWARE Y BASES DE DATOS, 8. **Proceedings…** Alicante: 2003. p. 603–612.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101: A Guide to Creating Your First Ontology**. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, Mar. 2001.

OGC. **OpenGIS® Geography Markup Language (GML) Implementation Specification. v. 3.1.0**. OGC 03-105r1. OpenGIS Consortium, 2004. Disponível em: <http://www.opengeospatial.org/standards/gml>. Acesso em: 15 jan. 2004.

OGC. **Open Geospatial Consortium, Inc.** Site oficial. Disponível em: <http://www.opengeospatial.org/>. Acessado em: 03 jan. 2007.

OLIVEIRA, C. M. de. **Manutenção da Nomenclatura de Logradouros e Numeração de Endereços, no Geoprocessamento de Belo Horizonte**. 53 f. 2001. Monografia (Especialização em Geoprocessamento) – Departamento de Cartografia, Universidade Federal de Minas Gerais, Belo Horizonte.

OLIVEIRA, P. A. de; OLIVEIRA, M. P. G. Usos de um Sistema de Informação Geográfica em Cadastro Técnico Municipal: a experiência de Belo Horizonte. **Informática Pública**, v. 7, n. 2, p. 67-84, 2005.

OLIVEIRA, J. L.; PIRES, F.; MEDEIROS, C. M. B. An Environment for Modeling and Design of Geographic Applications. **GeoInformatica**, v. 1, n. 1, p. 29-58, 1997.

OWL. **Web Ontology Language**. Especificação oficial da OWL. Disponível em: <http://www.w3.org/2004/OWL/>. Acessado em: 16 jan. 2007.

PARENT, C.; SPACCAPIETRA, S,; ZIMANYI, E. Spatio-Temporal Conceptual Models: Data Structures + Space + Time. In: INTERNATIONAL SYMPOSIUM ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS – ACM GIS, 7., 1999, Kansas City. **Proceedings…** Kansas City: ACM, 1999. p. 26-33.

PINHO, C. M. D. de; GOLTZ, E. **Construção de ontologias espaciais**: O lote urbano. Trabalho apresentado na disciplina Banco de Dados Geográficos. Instituto Nacional de Pesquisas Espaciais. São José dos Campos: DPI/INPE, 2003.

PROTÉGÉ. **What is protégé-owl?**. Site oficial do editor Protégé-OWL. Disponível em: <http://protege.stanford.edu/overview/protege-owl.html>. Acessado em: 16 jan. 2007.

RAHM, E.; BERNSTEIN, P. A. A survey of approaches to automatic schema matching. **The VLDB Journal**, n. 10, 2001, Springer-Verlag. p. 334-350.

RIZZO NETO, A.; REIS, L. C. DOS. **Manual de Quadras**. Cadastro Técnico Municipal. Belo Horizonte: PRODABEL, 1999.

RODRÍGUEZ, M. A.; EGENHOFER, M. J. Determining Semantic Similarity Among Entity Classes from Different Ontologies. **IEEE Transactions on Knowledge and Data Engineering**, v. 15, n. 2, Mar./Apr. 2003. p. 442-456.

RODRÍGUEZ, M. A.; EGENHOFER, M. J.; RUGG, R. D. Assessing Semantics Similarities Among Geospatial Feature Class Definitions. In: VCKOVSKI, A.; BRASSEL, K.; SCHEK, H. J. (eds.). **Interoperating Geographic Information**

Systems, INTEROP'99. Lecture Notes in Computer Science. v. 1580. Zurich: Springer-Verlang, 1999. p. 189-202.

ROSA, L. R. da. **Uma Ferramenta de Apoio a Determinação de Equivalências Semânticas entre Esquemas GML Utilizando Ontologias OWL**. 151 f. 2006. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Departamento de Informática e Estatística, Universidade Federal de Santa Catarina – UFSC, Florianópolis.

SALTON, G. **Dynamic Information and Library Processing**. New Jersey: Prentice Hall, 1975. 523 p.

SHETH, A. P.; KASHYAP, V. So Far (Schematically) yet So Near (Semantically). In: IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems, Lorne. **Proceedings...** Lorne: IFIP Transactions, 1992. p. 283-312

SILVA, R. A. B. e. **Interoperabilidade na representação de dados geográficos**: GeoBR e GML 3.0 no contexto da realidade dos dados geográficos no Brasil. 2003. 148 f. Dissertação (Mestrado em Computação Aplicada) - Pós-Graduação em Computação Aplicada, Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos.

SOTNYKOVA, A.; CULLOT, N.; VANGENOT, C. Spatio-Temporal schema integration with validation: a practical approach. In: MEERSMAN, R. *et al*. (eds.) **On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops**. Agia Napa, Cyprus: Springer, 2005. v. LNCS 3762, p. 1027-1036.

STOIMENOV, L.; STANIMIROVIC, A.; DJORDJEVIC-KAJAN, S. Realization of Component-Based GIS Application Framework. In: AGILE CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE, 7., 2004, Heraklion, Greece. **Proceedings…** Crete: Crete Unversity Press, 2004. p. 113-120.

THOMÉ, R. **Interoperabilidade em geoprocessamento: conversão entre modelos conceituais de sistemas de informação geográfica e comparação com o padrão OpenGIS**. 1998. 200 f. Dissertação (Mestrado em Computação Aplicada) – Ministério da Ciência e Tecnologia, Instituto Nacional de Pesquisas Espaciais – INPE, São José dos Campos.

UCGIS. **2002 Research Agenda**. University Consortium for Geographic Information Science, 2002. Disponível em: <http://http://www.ucgis.org/priorities/research/2002researchagenda.htm>. Acessado em: 23 mar. 2004.

XML. **Extensible Markup Language**. Disponível em: <http://www.w3.org/XML/>. Acessado em: 06 jan. 2007.

XML SCHEMA. **XML SCHEMA Part 2: Datatypes Second Edition**. Apresenta as definições de tipos de dados usados em XML Schema. 2004. Disponível em:<http://www.w3.org/TR/xmlschema-2/#datatype>. Acessado em: 16 já. 2007.

YUAN, M. Development of a Global Conceptual Schema for Interoperable Geographic Information. In: INTERNATIONAL CONFERENCE AND WORKSHOP ON INTEROPERATING GEOGRAPHIC INFORMATION SYSTEMS, 1997, Santa Barbara (EUA). **Proceedings…** Santa Barbara: NCGIA, 1997.

APÊNDICES

APÊNDICE I - GEOGRAPHY MARKUP LANGUAGE

Este apêndice apresenta um breve estudo da *Geography Markup Language* (GML) realizado com base nos trabalhos de Silva (2003) e Hess (2004) e na especificação atual da linguagem (OGC, 2004).

I.a Apresentação

A *Geography Markup Language* (GML) é uma gramática escrita em XML (*eXtensible Markup Language*) *Schema* para a modelagem, transporte e armazenamento de informação geográfica. Os conceitos chaves usados pela GML para modelar o mundo real são extraídos da especificação feita pelo *Open*GIS para modelagem de fenômenos geográficos e da série de normas ISO 19100. A GML provê uma variedade de objetos para descrever a geografia, incluindo feições (*features*), sistemas de referência de coordenadas, geometria, topologia, tempo, unidades de medida e valores gerais (OGC, 2004).

A GML foi projetada para suportar interoperabilidade, obtida através da provisão de *tags* geométricas básicas - ou seja, todos os sistemas que usam GML usam as mesmas *tags* -, um modelo de dados comum e um mecanismo para a criação e compartilhamento de esquemas de aplicação através do uso de *namespaces* XML (SILVA, 2003).

A primeira versão da GML, publicada em maio de 2000, teve como base a combinação de *Document Type Definition* (DTD) e *Resource Definition Framework* (RDF). Para a versão 2.0, publicada em março de 2001, a GML foi inteiramente baseada em XML *Schema*, enriquecendo sua representação. Esta mudança deu-se em função de a DTD ter caído em desuso (SILVA, 2003). A partir da versão 3.0 (OGC, 2004), publicada em janeiro de 2003, a GML busca consolidar-se como um padrão para o transporte e armazenamento de informação geográfica, incluindo propriedades espaciais e não espaciais das feições geográficas. Esta última versão traz muito mais recursos, elaborados a partir das necessidades atuais e de deficiências encontradas nas versões

anteriores, principalmente no que se refere à forma de representar objetos geográficos (SILVA, 2003).

I.b Representação de feições geográficas

Uma feição geográfica, ou *feature*, corresponde a uma unidade da informação geográfica. Na GML, ela é uma abstração de um fenômeno do mundo real que é associado a uma localização relativa na superfície da Terra. Assim, uma representação do mundo real pode ser vista como um conjunto de *features*. O estado de uma *feature* é definido por um conjunto de propriedades, que podem ser espaciais (geométricas) e descritivas (simples, como inteiros, *strings* etc.), sendo que cada propriedade é vista como uma tripla *{nome, tipo, valor}*. O número de propriedades que uma *feature* pode ter é determinado pela sua definição de tipo. *Features* geográficas são aquelas cujas propriedades podem ter valores espaciais (geométricos). Uma coleção de *features* pode, ela mesma, ser vista como uma *feature*. Como conseqüência, uma coleção de *features* tem um tipo e pode ter propriedades distintas próprias dela, além daquelas herdadas das *features* que a compõem (OGC, 2004).

Como exemplo, uma cidade pode ser representada como uma coleção de *features,* em que cada *feature* individual representa fenômenos como rios, estradas, colégios, hospitais etc. Cada um destes tipos de feições tem um nome e propriedades tipadas próprias. As propriedades podem ser de tipo simples (inteiro, *string*, ponto flutuante, *boolean* etc.) ou de tipos geométricos (ponto, linha e polígono). Assim, a *feature Rio* pode ter uma propriedade simples chamada *nome*, cujo valor deve ser do tipo *string*, e uma propriedade geométrica chamada *centerLineOf*, cujo valor deve ser do tipo *linha*. Da mesma forma que uma *feature* pode ter mais de uma propriedade simples, ela também pode ter mais de uma propriedade geométrica. Neste caso, uma *feature* chamada *Colegio*, além da propriedade *nome*, pode ter uma propriedade do tipo *ponto*, chamada *localização*, e uma propriedade do tipo *polígono*, chamada *campus* (OGC, 2004).

Até a versão 2, o modelo de dados da GML tinha capacidade de representar apenas feições simples, ou seja, *features* cujas propriedades geométricas são restritas a geometrias simples (ponto, linha e polígono), para as quais são definidas coordenadas

em até duas dimensões, não sendo possível representar superfícies, por exemplo (SILVA, 2003).

A partir da versão 3, a GML apresenta o conceito de objetos geográficos, ou seja, uma feição geográfica é uma especialização de um objeto geográfico (SILVA, 2003). Além disso, foram adicionadas novas características (OGC, 2004):

- Maior representação de fenômenos geo-espaciais, incluindo *features* com geometria 3D complexa e não-linear, *features* com topologia 2D, *features* com propriedades temporais, *features* dinâmicas, *coverages*[2] e *observations* (observações)[3];
- Provê suporte mais explícito para propriedades de *features* e outros objetos cujos valores são complexos;
- Provê suporte para representação espacial, sistemas de referência temporal, unidades de medidas e informações padrões;
- Possibilita o uso de sistemas de referência, unidades e informações padrões na representação de fenômenos geo-espaciais, observações e valores;
- Possui estilos de representação padrões para visualização de *features* e *coverages*;
- Está em conformidade com os padrões da série ISO 19100.

I.c Classes de objetos GML

A GML fornece uma variedade de tipos de objetos (classes) usados para descrever:

- *Features* (incluindo *coverages* e *observations*);
- Sistemas de referência de coordenadas;
- Unidades de medidas;
- Valores (usados como propriedades de uma *feature*);

---

[2] *Coverage* é um tipo de *feature* que tem uma função que retorna valores dentro de sua área de cobertura a partir de qualquer posição dentro de seu domínio espaço-temporal (OGC, 2004).

[3] Uma *observação* modela a ação de observar, que pode ser com uma câmera, uma pessoa ou outra forma ou instrumento. É uma ação de reconhecer e anotar um fato ou ocorrência, muitas vezes envolvendo medição com instrumentos. Uma observação é considerada uma *feature* GML com um tempo no qual a observação é realizada e com um valor para a observação.

- Topologia e geometria (usados como propriedades de uma *feature*);
- Suporte temporal (usado como propriedade de uma *feature*).

Estas classes são definidas em uma hierarquia, como mostra o diagrama UML (*Unified Modeling Language*) na Figura 33.

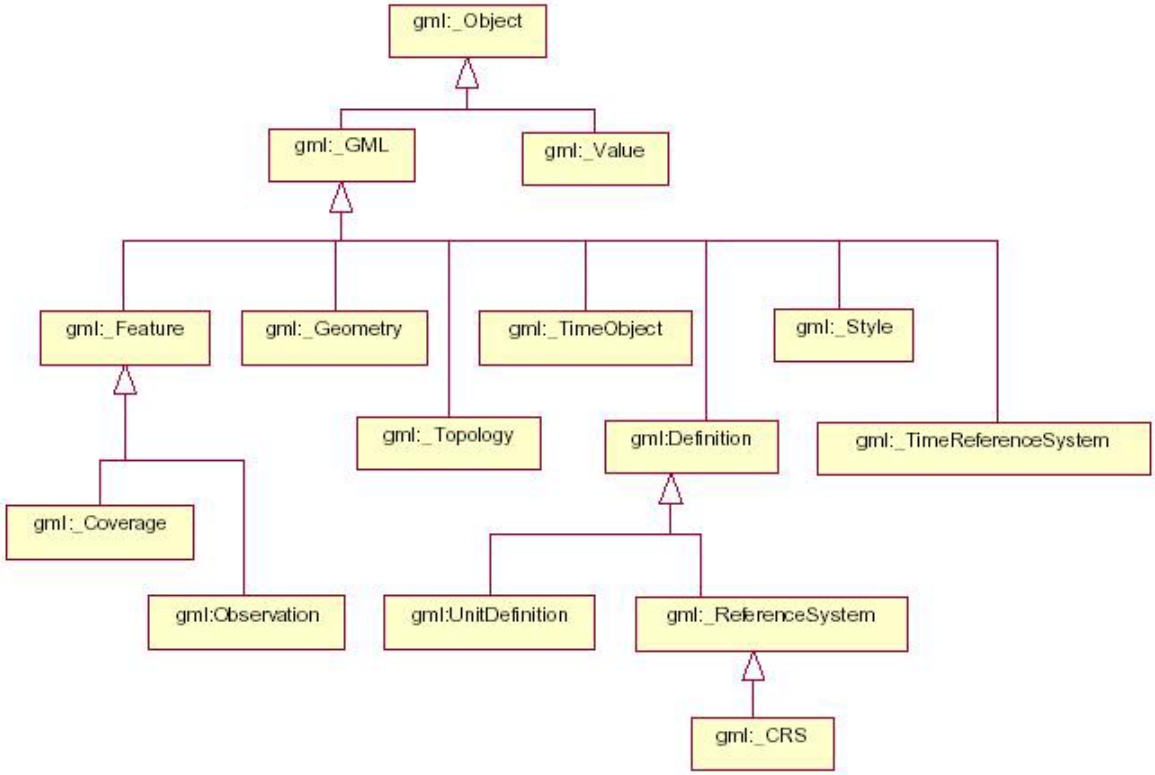


FIGURA 33: Hierarquia de classes GML
(FONTE: OGC, 2004)

A expansão da versão 3.x da GML, adicionando estas características, refletiu no aumento do número de esquemas base (33 esquemas no total) em relação às versões anteriores. Entretanto, segundo a sua especificação (OGC, 2004), poucas aplicações usam todas as definições presentes na GML. Assim, desenvolvedores podem usar um subconjunto selecionado de esquemas suficiente, para suportar as definições em seus esquemas de aplicação.

I.d Esquemas Base da GML

A hierarquia de dependência entre esquemas base da GML é representada na Figura 34 (OGC, 2004).

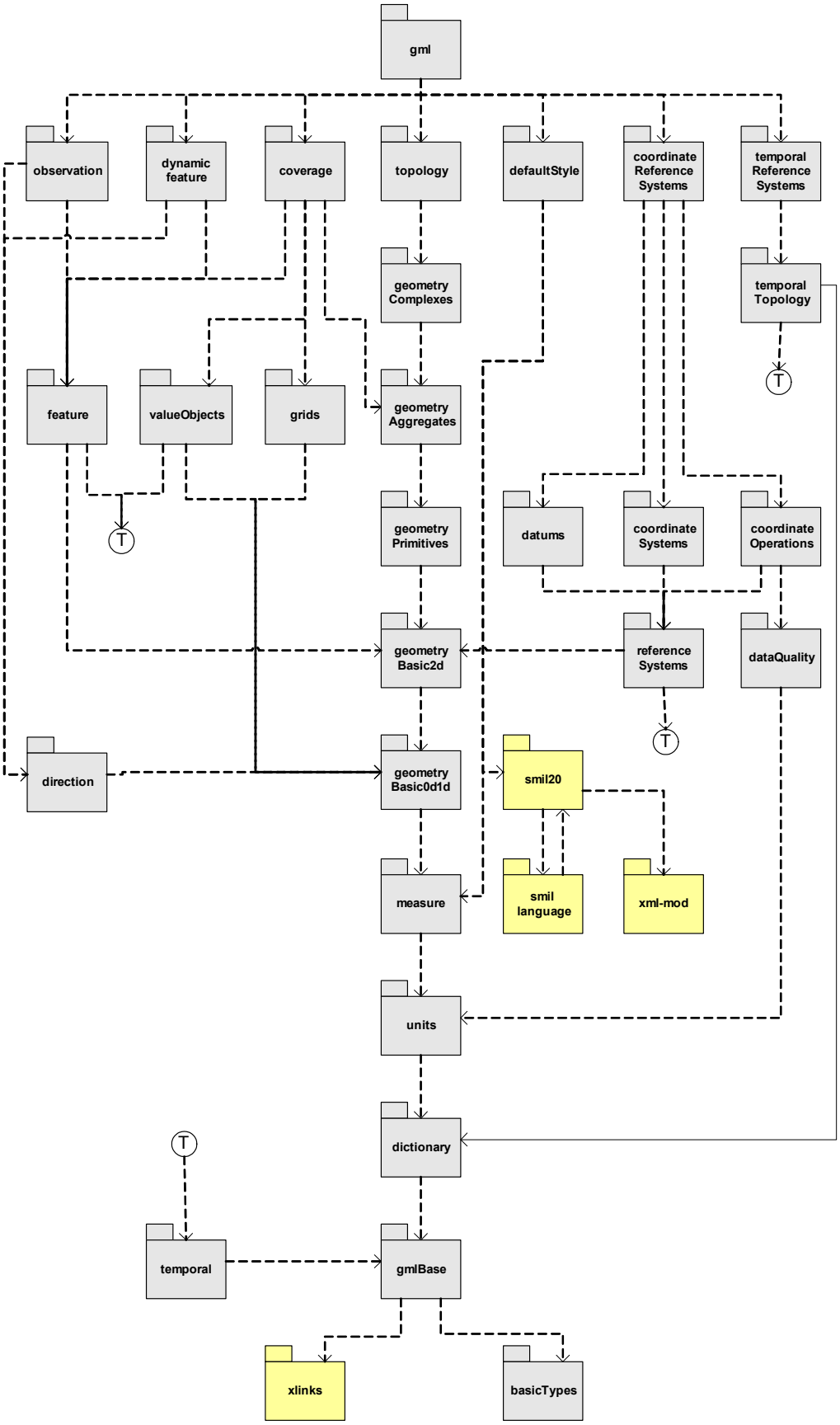


FIGURA 34: Dependência entre esquemas base GML
(FONTE: adaptado de OGC, 2004)

Observa-se que desta hierarquia podem ser retirados sete esquemas independentes dos demais, o que reforça a idéia de se usar apenas aqueles esquemas base necessários para a aplicação do usuário:

- *observation.xsd*
- *dynamicFeature.xsd*
- *coverage.xsd*
- *topology.xsd*
- *defaultStyle.xsd*
- *coordinateReferenceSystems.xsd*
- *temporalReferenceSystems.xsd*

O Quadro 15 apresenta uma descrição dos esquemas base da GML.

QUADRO 15: Descrição dos esquemas base da GML

Esquema base GML	Descrição
basicTypes.xsd	Define tipos simples e genéricos para representação.
coordinateOperations.xsd	Define elementos e tipos para codificar em XML definições de operações de coordenadas, incluindo transformações e conversões e outros subtipos específicos de operações. Usado em conjunto com o *referenceSystems.xsd*.
coordinateReferenceSystems.xsd	Define elementos e tipos para codificar em XML definições de sistemas de referência de coordenadas, incluindo subtipos específicos de sistemas de referência de coordenadas. Usado em conjunto com o *referenceSystems.xsd*.
coordinateSystems.xsd	Define elementos e tipos para codificar em XML definições para eixos de sistemas de coordenadas espaço-temporais, incluindo subtipos específicos de sistemas de coordenadas. Usado em conjunto com o *referenceSystems.xsd*.
coverage.xsd	Define os tipos que representam campos geográficos (*coverages*), para representar fenômenos contínuos do mundo real.
dataQuality.xsd	Define elementos e tipos para codificar em XML informações da qualidade dos dados de localização, necessários para descrever a exatidão da localização de operações de coordenadas. Usado em conjunto com o *units.xsd*.
datums.xsd	Define elementos e tipos para codificar em XML definições de *datums* e outros tipos específicos de *datums* espaço-temporais. Usado em conjunto com o *referenceSystems.xsd*.
defaultStyle.xsd	Define elementos usados na separação entre dados e apresentação.
dictionary.xsd	Permite a definição de novos conceitos em GML necessários para a aplicação. Não define os elementos propriamente ditos, mas sim seus significados e descrições.
direction.xsd	Provê um elemento de propriedade padrão para descrever direção e objetos associados que podem ser usados para expressar orientação, direção, rumo, comportamento ou outros aspectos direcionais de *features* geográficos.
dynamicFeature.xsd	Define um tipo especial de elemento temporal, utilizado para monitorar o comportamento e registros históricos de eventos de *features* que tenham propriedades que alteram seu estado no tempo.

Esquema base GML	Descrição
feature.xsd	É a base para a codificação das *features* de interesse da modelagem, sejam eles geográficos (espaciais) ou não. Todo objeto representado em GML é derivado deste esquema.
geometryAggregates.xsd	Geometrias agregadas são agregações arbitrárias de elementos geográficos. Não assumem ter qualquer estrutura interna adicional e são usados para juntar pedaços de geometria de um tipo específico.
geometryBasic0d1d.xsd	Descreve o modelo geométrico básico usado na GML (primitivas geométricas simples de dimensões 0D e 1D).
geometryBasic2d.xsd	Descreve o modelo geométrico básico usado na GML (primitivas geométricas simples de dimensão 2D).
geometryComplexes.xsd	Geometrias complexas são coleções fechadas de primitivas geométricas, isto é, elas contém seus limites.
geometryPrimitives.xsd	Especifica primitivas geométricas adicionais para descrever situações do mundo real que requerem um modelo geométrico mais expressivo.
gml.xsd	Esquema raiz da GML (*top level*).
gmlBase.xsd	Primeiro e mais básico esquema da GML 3.x. Define os tipos básicos da GML e os elementos que servem de suporte aos metadados. Serve como base para a definição de outros tipos (objetos ou propriedades).
grids.xsd	Provê as estruturas geométricas para *grid*, usadas na descrição de regiões em grade e outras aplicações.
measures.xsd	Esquema usado para especificar definições de tipos de medidas como volume, comprimento, peso, ângulo, fator de escala, tempo, área, velocidade e tamanho de *grid* etc. Alguns já estão pré-definidos, mas é possível definir novos. Estende os esquemas *units.xsd* e *basicTypes.xsd*.
observation.xsd	Modela uma *feature* de observação, ou seja, descreve os metadados associados com um evento de captura de informação, juntamente com um valor para o resultado da observação.
referenceSystems.xsd	Define elementos e tipos para codificar em XML definições dos tipos de objetos GML usados para sistemas de referência de coordenadas.
temporal.xsd	Define o suporte temporal da GML, para uso quando o estado das *features* modeladas muda com o passar do tempo.
temporalReferenceSystems.xsd	Provê construtores para tratamento de vários estilos de sistemas de referência temporal.
temporalTopology.xsd	Provê construtores para tratamento de topologias complexas e relacionamentos entre *features* temporais. As características geométricas temporais são representadas como instantes e períodos.
topology.xsd	Define o esquema topológico da GML, permitindo definir alguns tipos de relacionamentos espaciais entre objetos.
units.xsd	Esquema utilizado para definir unidades de medidas utilizadas na aplicação. Algumas já estão pré-definidas, mas é possível definir novas. Usado em conjunto com o *gmlBase.xsd*.
valueObjects.xsd	Descreve elementos e tipos para valores genéricos.
xlink.xsd	Este esquema é uma implementação do OGC (*Open Geospatial Consortium*) para a especificação *XLink* usando XML *Schema*. Ela pode ser alterada em algum *release* futuro por um esquema equivalente do W3C (*World Wide Web Consortium*). Usado para referenciar elementos externos ao documento GML.
smil20.xsd	Esquema adicional, usado para referenciar elementos externos ao documento GML.
smil20-language.xsd	Esquema adicional, usado para referenciar elementos externos ao documento GML.
xml-mod.xsd	Esquema adicional, usado para referenciar elementos externos ao documento GML.

(FONTE: OGC, 2004; HESS, 2004)

APÊNDICE II - MODELO CONCEITUAL OMT-G

Este apêndice apresenta um breve estudo sobre o modelo conceitual OMT-G (*Object Modeling Technique for Geographic Applications*), proposto por Borges (1997) e apresentado em Borges, Davis Junior e Laender (2001). A adoção do modelo conceitual OMT-G dá-se por dois motivos: primeiro, por este ter amplo uso na literatura e, segundo, por se enquadrar nas necessidades do trabalho, uma vez que o esquema GML (*Geography Markup Language*) criado baseou-se no trabalho de Bertini (2003), que utiliza o OMT-G. Além disso, outras referências citadas, como Pinho e Goltz (2003), são baseadas em trabalhos desenvolvidos na PRODABEL - Empresa de Informática e Informação do Município de Belo Horizonte, que utiliza amplamente o OMT-G e que tem influência na origem do modelo.

II.a Introdução

Um modelo de dados é um conjunto de conceitos usados para descrever a estrutura e as operações em um banco de dados (ELMASRI e NAVATHE, 2004). O modelo de dados busca sistematizar o entendimento que é desenvolvido a respeito de objetos e fenômenos que são representados em um sistema computadorizado (BORGES, DAVIS JÚNIOR e LAENDER, 2005).

Modelos de dados tradicionais (como ER, OMT e UML) não são adequados para a representação de dados geográficos, uma vez que não possuem primitivas apropriadas para sua modelagem. Dados geográficos possuem aspectos peculiares, particularmente com respeito à codificação da localização espacial e do tempo de observação, além do registro de fatores externos, como a sua precisão de obtenção (BORGES, DAVIS JÚNIOR e LAENDER, 2005). Assim, verifica-se a necessidade de modelos conceituais próprios para representar o espaço geográfico.

Os primeiros modelos de dados para aplicações geográficas eram específicos para as estruturas internas dos Sistemas de Informações Geográficas (SIGs), o que

limitava seu poder de expressão às estruturas disponíveis no SIG utilizado pelo usuário (BORGES, DAVIS JÚNIOR e LAENDER, 2005).

Para representar a realidade dos dados geográficos de forma mais próxima ao modelo mental do usuário, um modelo conceitual deve considerar diversos fatores, dentre eles (BORGES, DAVIS JÚNIOR e LAENDER, 2005):

- Transcrição da informação geográfica em unidades lógicas de dados;
- Forma como as pessoas percebem o espaço;
- Natureza diversificada dos dados geográficos;
- Existência de relações espaciais, como topológicas, métricas, de ordem e *fuzzy*.

Estão disponíveis diversos modelos de dados voltados para aplicações geográficas, como MODUL-R (BÉDARD *et al.*, 1996), GeoOOA (KÖSTERS, PAGEL e SIX, 1997), GMOD (OLIVEIRA, PIRES e MEDEIROS, 1997), *GeoFrame* (LISBOA FILHO, 1997), MADS (PARENT, SPACCAPIETRA e ZIMANYI, 1999) e OMT-G (BORGES, 1997). Alguns deles estendem modelos tradicionais, enquanto outros são totalmente novos.

Para a escolha de qual modelo adotar, de acordo com Borges, Davis Junior e Laender (2001), deve-se observar as necessidades de modelagem quanto à abstração de conceitos geográficos, ao atendimento de requisitos usuais para modelos de dados e à possibilidade de mapeamento dos esquemas produzidos para a implementação em Sistemas Gerenciadores de Banco de Dados (SGBDs) espaciais.

O modelo OMT-G introduz primitivas geográficas ao conjunto de primitivas definidas para o diagrama de classes da *Unified Modeling Language* (UML), procurando aumentar sua capacidade de representação semântica (BORGES, DAVIS JÚNIOR e LAENDER, 2005). O fato de ser baseado em diagramas UML também justifica a adoção do modelo OMT-G, uma vez que a UML vem sendo largamente usada no projeto de sistemas em geral, o que facilita o aprendizado do modelo. Além disso, a própria especificação da *Geography Markup Language* (GML) apresenta diversos diagramas na notação adotada pela UML.

Para o desenvolvimento de aplicações geográficas, o modelo OMT-G propõe o uso de três diferentes diagramas:

- *Diagrama de classes*: especifica todas as classes, suas representações e relacionamentos;
- *Diagrama de transformação*: especifica o processo de múltiplas representações de uma classe ou como deve ser feita a derivação de uma classe a partir de outra;
- *Diagrama de apresentação*: especifica as alternativas de visualização que cada representação pode adquirir.

Neste trabalho, utiliza-se o diagrama de classes do modelo OMT-G, uma vez que este diagrama é usado para descrever a estrutura e o conteúdo de um banco de dados geográfico. Usado no nível de representação conceitual, o diagrama de classes contém regras e descrições que definem conceitualmente como os dados são estruturados e que tipo de representação é atribuído para cada classe. A notação usada pelo modelo OMT-G é semelhante à usada no diagrama de classes da UML, apenas acrescentando no canto superior esquerdo das classes georreferenciadas um espaço usado para indicar a forma geométrica da representação (Figura 35).

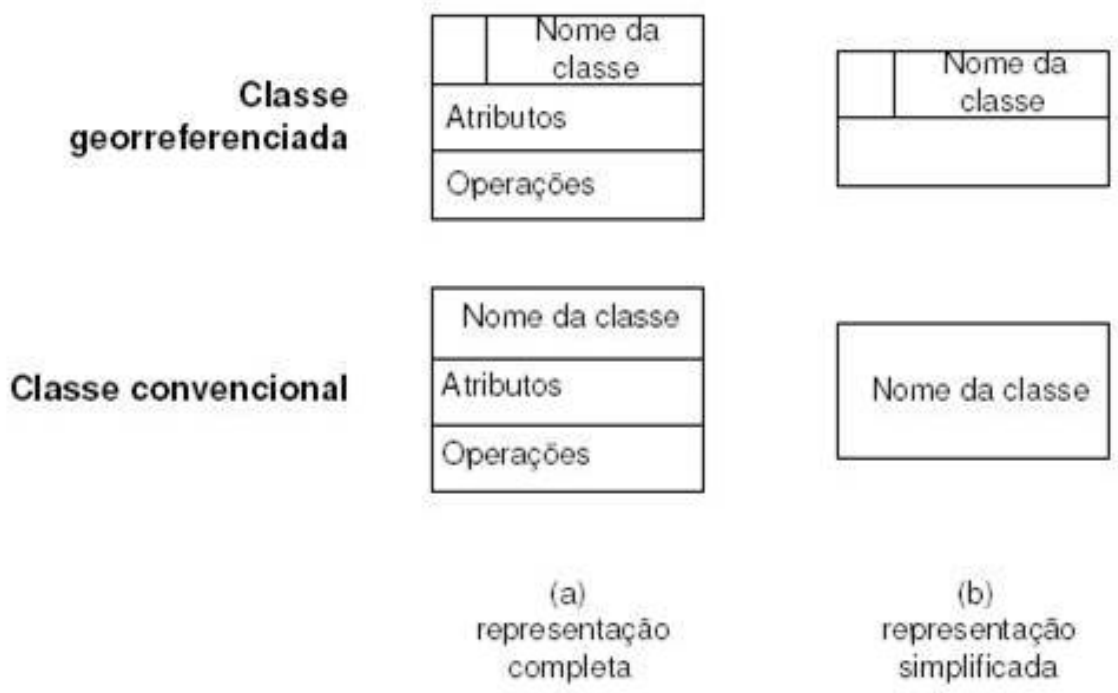


FIGURA 35: Notação gráfica para as classes do modelo OMT-G
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

As primitivas para o diagrama de classes são (BORGES, DAVIS JÚNIOR e LAENDER, 2005):

- *Classes*: as classes definidas pelo modelo OMT-G representam os três grupos de dados encontrados em aplicações geográficas (contínuos, discretos e não-espacias) e podem ser georrefenciadas ou convencionais. Classes georreferenciadas descrevem objetos que possuem representação espacial e estão associadas a regiões da superfície da terra, representando as visões de campo (Figura 36) e de objeto (Figura 37).

FIGURA 36: Exemplo de notação para a visão de campo
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

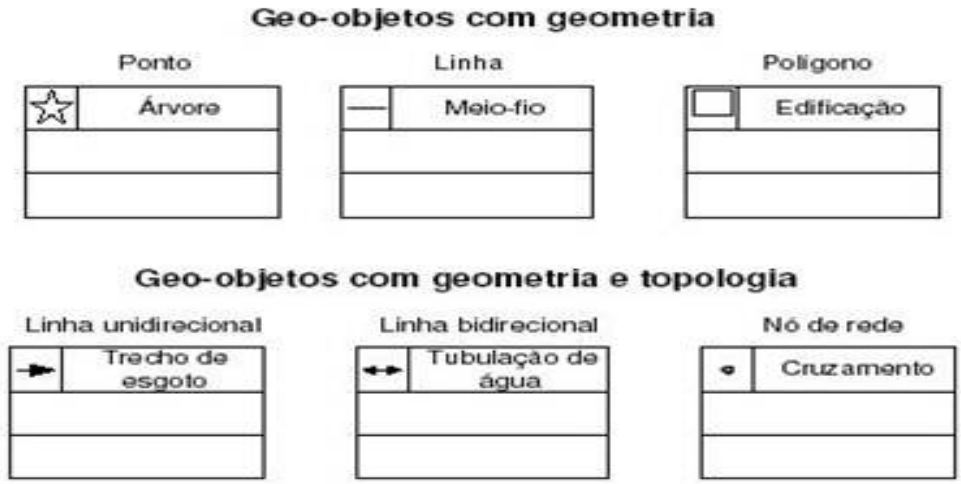


FIGURA 37: Exemplo de notação para a visão de objeto
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

- *Relacionamentos*: o modelo OMT-G representa três tipos de relacionamentos entre suas classes: associações simples, relacionamentos topológicos em rede e relacionamentos espaciais. Associações simples representam relacionamentos estruturais entre objetos de classes diferentes, convencionais ou georreferenciadas, e são indicadas por linhas contínuas (Figura 38a). Os relacionamentos de rede representam objetos conectados uns com os outros, no formato arco-nó, e são indicados por duas linhas pontilhadas paralelas (Figura 38c). O modelo prevê, ainda, a existência de

estruturas de rede sem nós (Figura 38d). Relacionamentos espaciais representam relações topológicas, métricas, de ordem e *fuzzy*, e são indicados por linhas pontilhadas (Figura 38b).

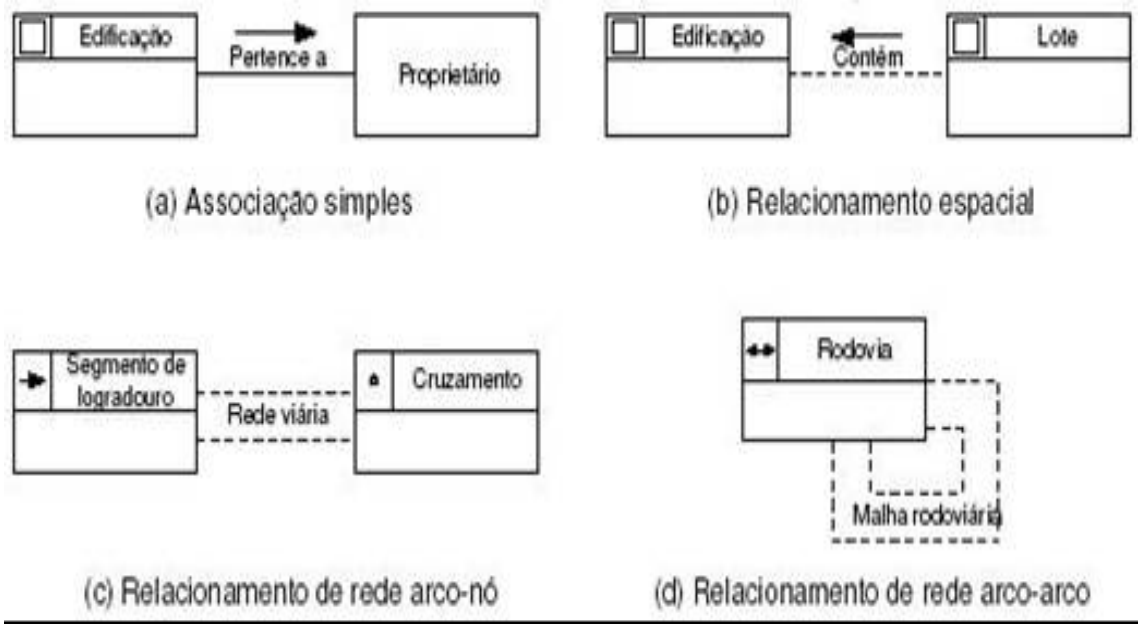


FIGURA 38: Exemplo de notação de relacionamentos
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

- *Cardinalidade*: representa o número de instâncias de uma classe que podem estar associadas a instâncias de outra classe. A notação de cardinalidade é demonstrada na Figura 39.

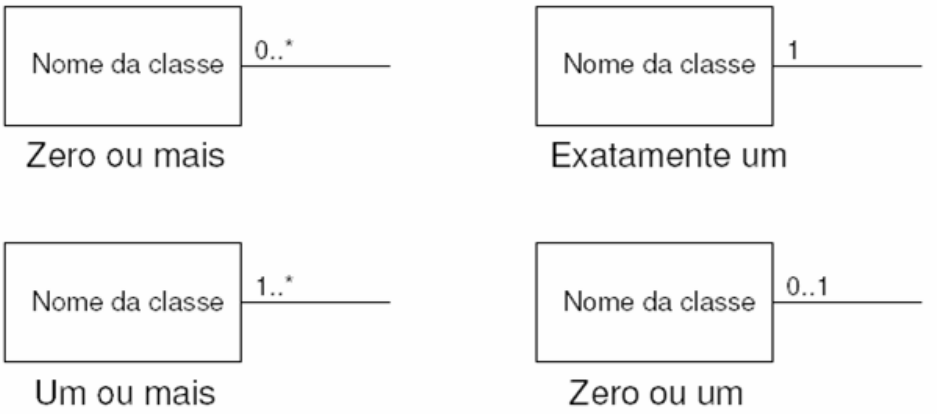


FIGURA 39: Exemplo de notação de cardinalidade
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

- *Generalização* e *especialização*: generalização é o processo de definição de classes mais genéricas (superclasses) a partir de classes com características

semelhantes (subclasses). Especialização é o processo pelo qual classes mais específicas são detalhadas a partir de classes genéricas, herdando atributos, métodos e associações da superclasse e adicionando novos atributos e métodos (Figuras 40 e 41).

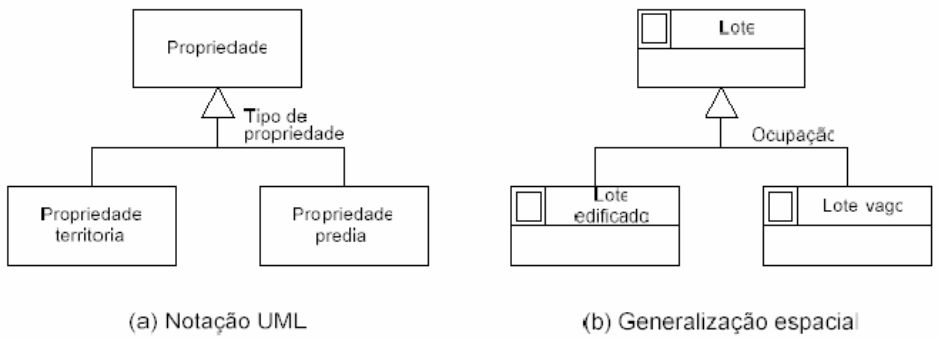


FIGURA 40: Exemplo de notação para generalização e especialização
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

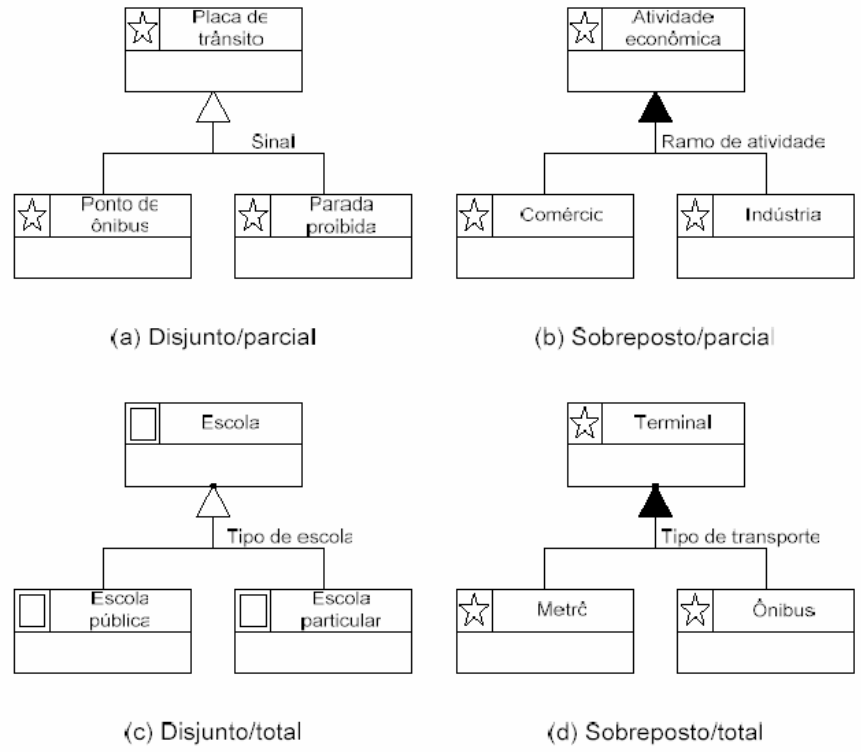


FIGURA 41: Exemplo de generalização espacial
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

- *Agregação*: considera que um objeto é formado a partir de outros objetos (Figura 42). Quando a agregação ocorre entre classes georreferenciadas é necessário usar a agregação espacial (Figura 43).

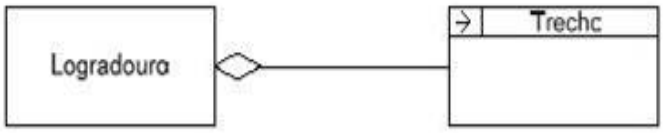


FIGURA 42: Exemplo de agregação entre classe convencional e classe georreferenciada
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

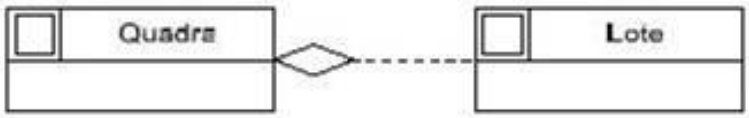
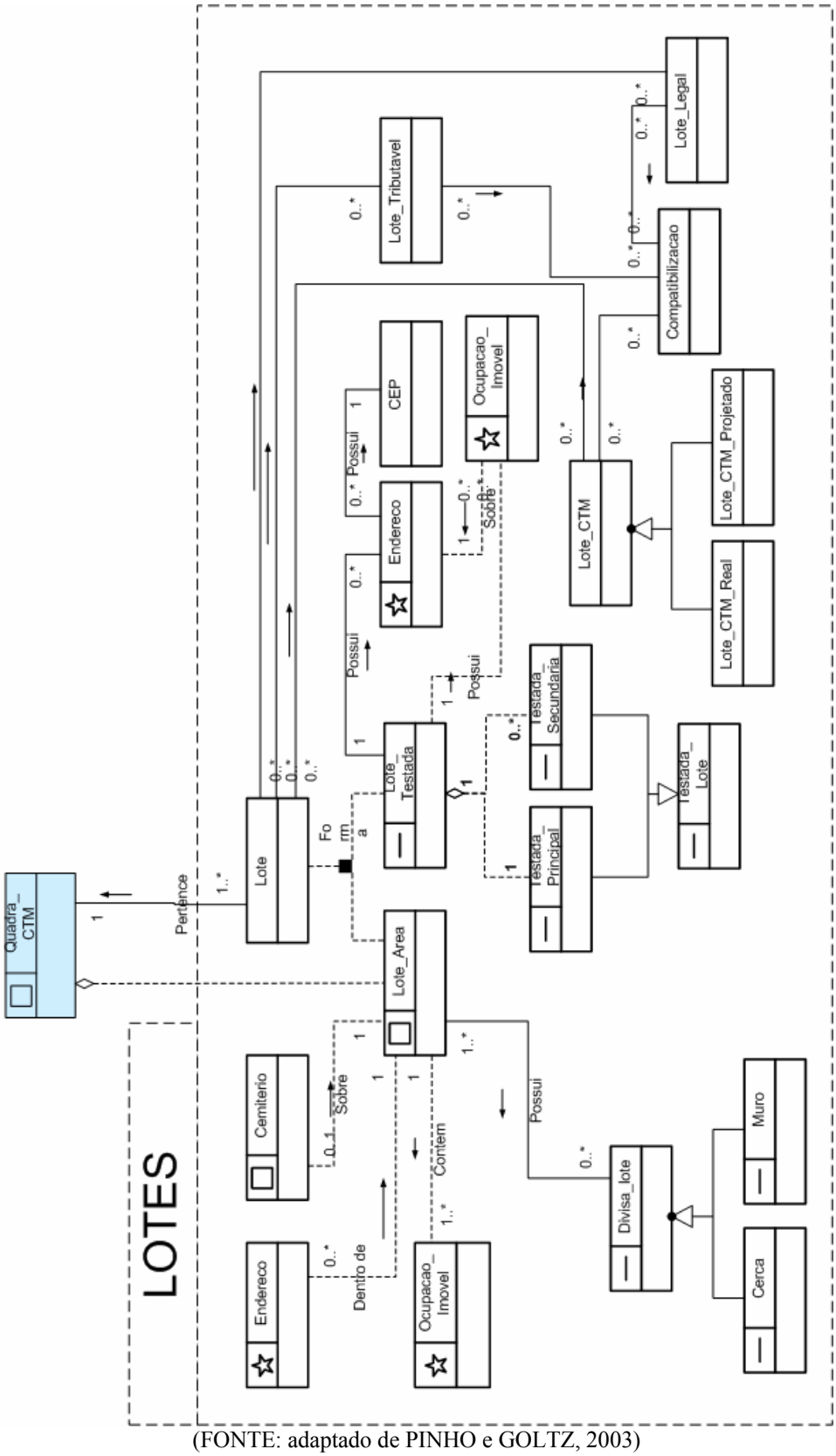


FIGURA 43: Exemplo de agregação espacial
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

- *Generalização conceitual* ou *cartográfica*: indica transformações na representação da informação espacial, em que um objeto do mundo real pode ter diversas representações espaciais. Por exemplo, no mapa de uma cidade (escala pequena), uma construção (como uma igreja) pode ser representada como um ponto, enquanto que no mapa de um bairro específico (escala maior), esta construção pode ser representada como um polígono. A generalização conceitual pode ser de acordo com a forma geográfica (classes sobrepostas) (Figura 44a) ou de acordo com a escala (classes disjuntas) (Figura 44b).

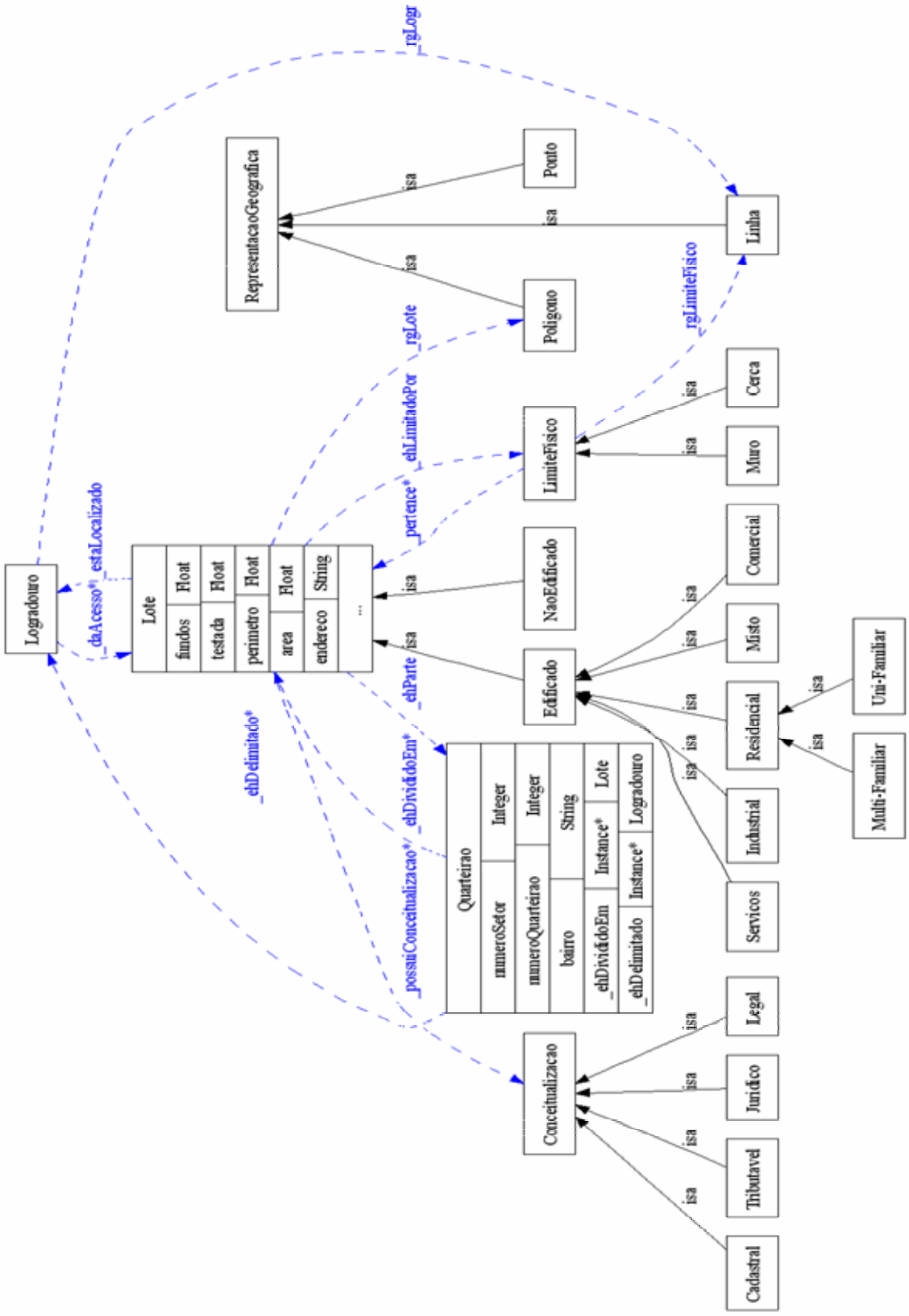
FIGURA 44: Exemplo de notação para a generalização conceitual
(FONTE: BORGES, DAVIS JÚNIOR e LAENDER, 2005)

APÊNDICE III - ESQUEMA OMT-G DO PACOTE LOTES



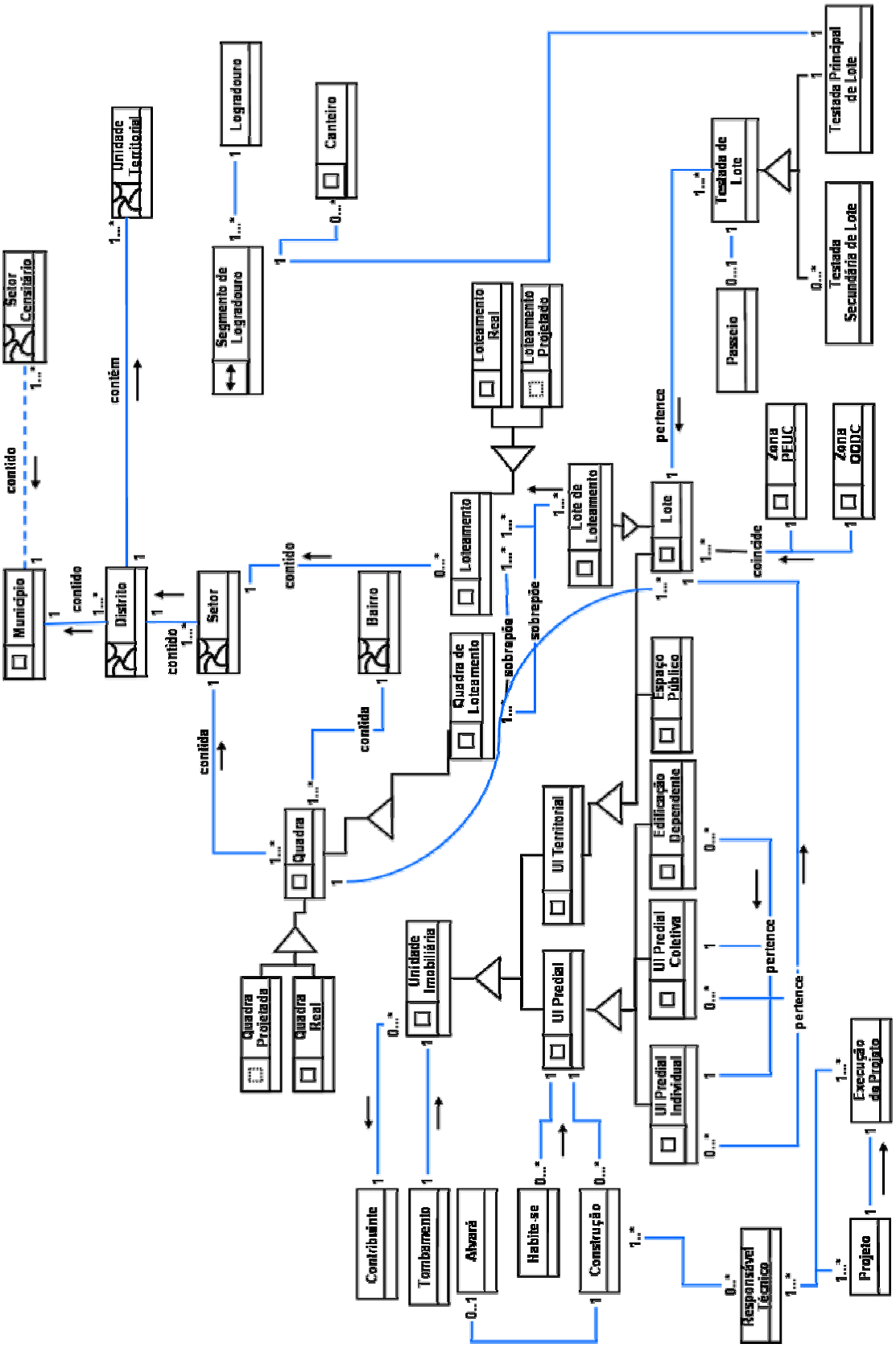
(FONTE: adaptado de PINHO e GOLTZ, 2003)

APÊNDICE IV - ONTOLOGIA DO DOMÍNIO LOTE URBANO



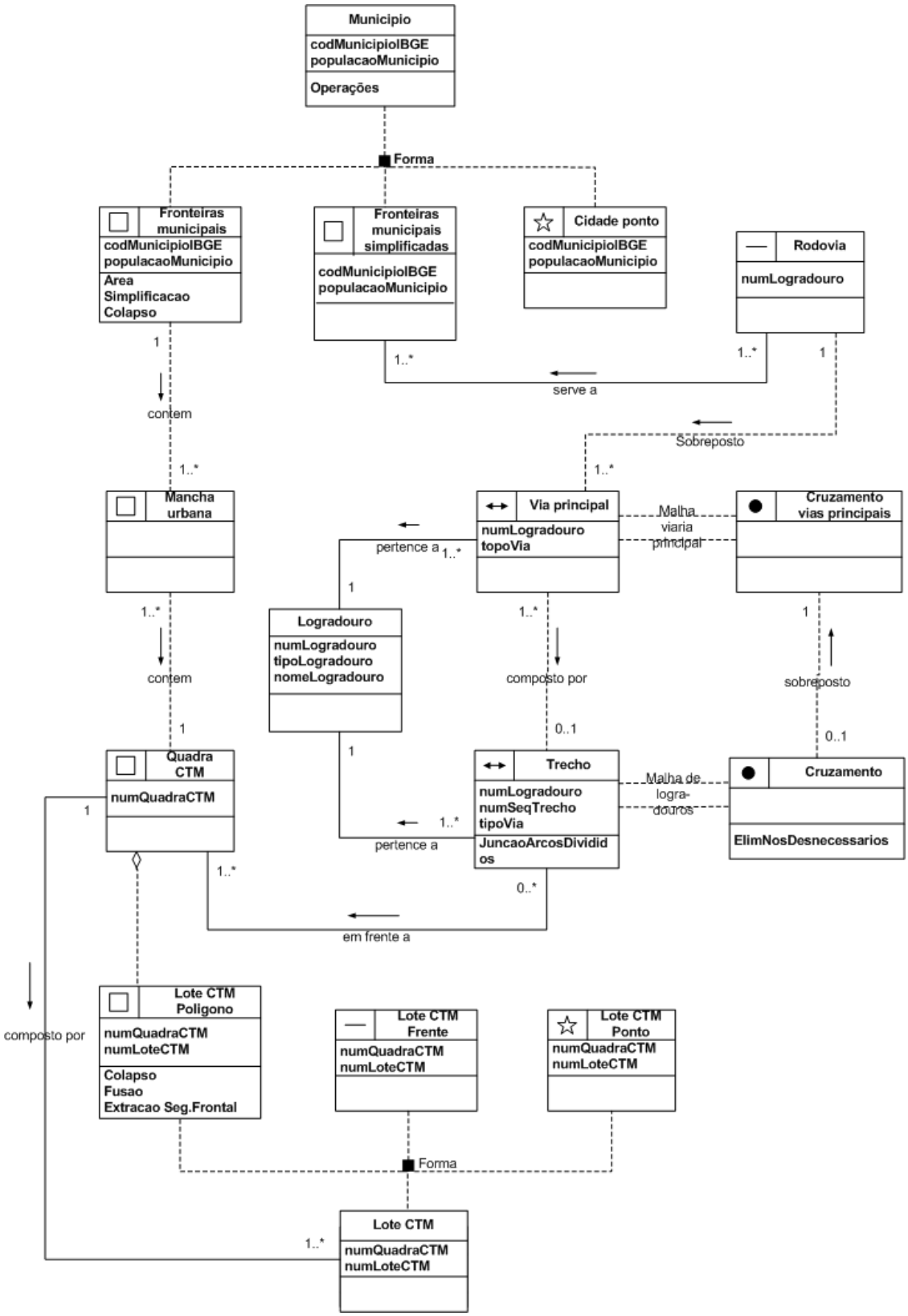
(FONTE: criado a partir de BERTINI, 2003)

APÊNDICE V - ESQUEMA CONCEITUAL DO SIG LAGES



(FONTE: adaptado do esquema conceitual fornecido pela Prefeitura do Município de Lages)

APÊNDICE VI - EXEMPLO DE ESQUEMA CONCEITUAL OMT-G



(FONTE: adaptado de BORGES, DAVIS JÚNIOR e LAENDER, 2005, p. 131)

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo