VIII Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, November 19-22, 2006, INPE, p. 283-294.

Um Método para Determinar a Equivalência Semântica entre Esquemas GML

Angelo Augusto Frozza[1,2], Ronaldo dos Santos Mello[2]

[1]Departamento de Ciências Exatas e Tecnológicas
Universidade do Planalto Catarinense (UNIPLAC)
Caixa Postal 525 88.509-900 Lages SC Brasil

[2]Departamento de Informática e de Estatística
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 88.049-900 Florianópolis SC Brasil

{frozza,ronaldo}@inf.ufsc.br

Abstract. *One of the difficulties faced by Geographic Information Systems (GIS) is exchanging information among distinct systems. The Geography Markup Language (GML) specifies a set of rules for geographic information transport and storage. However, different GIS can use different GML schemas, generating interoperability problems at the semantic level. This paper proposes a semi-automatic method for determining semantic equivalence among heterogeneous GML schemas for the urban registration domain, using an ontology as a semantic reference.*

Resumo. *Uma das dificuldades enfrentadas por Sistemas de Informações Geográficas (SIG) é a troca de informações entre sistemas distintos. A Geography Markup Language (GML) define um conjunto de regras para o transporte e armazenamento de informação geográfica. Porém, SIGs diferentes podem usar esquemas GML diferentes, gerando problemas de interoperabilidade em nível semântico. Este trabalho propõe um método semi-automático para determinar a equivalência semântica entre esquemas GML heterogêneos no domínio de cadastro urbano, usando uma ontologia como uma referência semântica.*

1. Introdução

A complexidade e a riqueza dos modelos de dados geográficos encontrados em Sistemas de Informações Geográficas (SIGs) proporcionam diferentes formas para representar a realidade geográfica. A ampliação e a diversificação do uso de SIGs nas organizações gerou a necessidade de intercâmbio de informações georeferenciadas entre fontes autônomas e heterogêneas [Zhang *et al.* 2004].

A troca de informações entre SIGs heterogêneos apresenta incompatibilidades nos níveis sintático e semântico. O nível sintático refere-se ao esquema utilizado em cada sistema para armazenamento e documentação dos dados. A resolução de conflitos neste nível baseia-se na conversão sintática direta de formatos de exportação e importação.

Todavia, a simples transferência e re-formatação de dados de um sistema para outro não garante que os dados tenham significado para o novo usuário. A interoperabilidade entre SIGs requer uma interpretação semântica para explicar a correspondência dos conceitos entre diferentes sistemas [Câmara *et al.* 1999].

Soluções visando à interoperabilidade semântica entre SIGs apontam como tendência o uso de padrões como a *Geography Markup Language* (GML) [OGC 2003]. O objetivo da GML é oferecer um conjunto de construções básicas, incluindo o modelo de *features* geográficas e uma coleção de meta-classes de objetos geográficos, com os quais um usuário pode estruturar e descrever seus dados georeferenciados.

Apesar dos avanços encontrados na versão atual da GML (3.1.1), a representação semântica ainda é limitada. Para solucionar esse problema, pode-se associar alguma descrição ontológica aos esquemas GML para promover a interoperabilidade semântica entre SIGs. A OWL (*Web Ontology Language*) é a especificação mais recente do W3C (*World Wide Web Consortium*) para representação de ontologias e é compatível com a arquitetura da *Web* em geral e da *Web* Semântica em particular [OWL 2006].

Este artigo apresenta um método para a determinação semi-automatizada de equivalências semânticas entre esquemas GML distintos, usando uma ontologia como base de conhecimento comum. Sua contribuição encontra-se no apoio ao desenvolvimento de *softwares* que possibilitem a troca de informações entre bases de dados geográficos, garantindo a semântica dos dados. Como estudo de caso, utiliza-se o domínio de cadastro urbano, escolhido por ser pouco explorado em trabalhos relacionados e pelo grande potencial de aplicação prática.

Trabalhos relacionados [Brauner, Casanova e Lucena 2004; Morocho, Pérez-Vidal e Saltor 2003] abordam a interoperabilidade semântica entre SIGs em ambientes fortemente acoplados, com ênfase na transformação de consultas. Diferente destes, este trabalho enfatiza a integração de dados geográficos em um domínio de aplicação específico, considerando que as fontes de dados não estão interligadas, mas podem, regularmente, realizar trocas de dados geográficos para atualização de uma base comum.

Este artigo está organizado da seguinte forma: a seção 2 propõe o método para determinação de equivalências semânticas entre esquemas GML. As seções 3 a 5 descrevem em detalhes as três partes do método: pré-processamento, determinação de equivalências e catalogação do mapeamento, respectivamente. A seção 6 apresenta as considerações finais.

2. O Método de Determinação de Equivalência Semântica

O método proposto (Figura 1) descobre equivalências semânticas entre dois esquemas GML distintos: um representando os dados de um SIG principal (GML principal ou GML') e outro representando os dados importados de um segundo SIG (GML importado ou GML''). Os conceitos geográficos considerados pelo método são representados por uma ontologia (Apêndice), que é também utilizada na determinação das equivalências semânticas. A atualização da ontologia a partir de conceitos novos será tratada em trabalhos futuros.

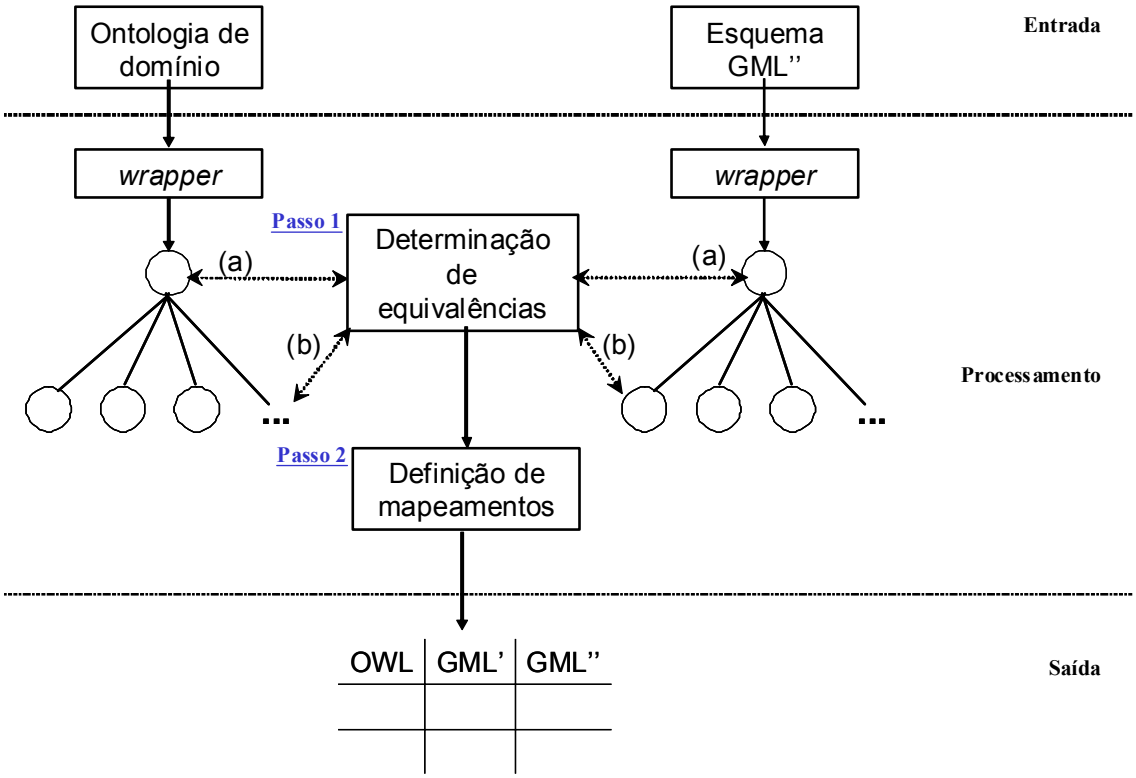


Figura 1. Método de determinação de equivalência semântica entre esquemas GML.

A execução do método segue dois passos:

1. Determinação da equivalência semântica entre o GML importado e a ontologia de domínio;
2. Definição e catalogação do mapeamento das equivalências entre o GML importado e o GML principal.

Certos elementos de um esquema GML importado são mapeados para certos elementos de um esquema GML principal. Duas entradas são analisadas: uma ontologia de domínio e um esquema GML importado qualquer. O esquema GML principal não faz parte da entrada, pois se assume que o método é aplicado no ambiente do SIG principal e que um mapeamento entre os conceitos da ontologia e os conceitos equivalentes do esquema GML principal já foi definido. Este mapeamento prévio garante a semântica dos conceitos do esquema GML principal.

Dois elementos de esquemas GML distintos são ditos semanticamente correspondentes quando possuem um determinado grau de similaridade. Este grau é determinado por métricas que consideram a semelhança entre os identificadores dos elementos, seus atributos e seus relacionamentos.

As métricas de similaridade aplicadas pelo método são uma adaptação das métricas encontradas em Dorneles *et al*. (2004):

- *Métricas para Valores Atômicos* (MAV): aplicadas a dados simples (elemento simples "*b*" na Figura 1), como *strings*. São dependentes do domínio da aplicação, ou seja, consideram as características dos dados da aplicação;

- *Métricas para Valores Complexos* (MCV): aplicadas à estrutura dos dados (elemento complexo "*a*" na Figura 1). Podem ser distintamente aplicadas a conjuntos (*tuplas*) ou coleções de valores.

Este conjunto de métricas foi escolhido por apresentar uma taxonomia adequada ao tratamento de dados XML. No caso, um elemento XML é tratado como uma árvore, considerando que pode ser 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.

Uma vez determinadas as equivalências semânticas entre a ontologia e o esquema GML importado, o método retorna como saída uma tabela de mapeamento na qual elementos do esquema GML importado correspondem semanticamente a elementos no esquema GML principal. O núcleo do método define correspondências entre conceitos da OWL e da GML (Tabela 1). Nestas correspondências, os conceitos associados possuem a mesma intenção em ambas as linguagens.

Tabela 1: Mapeamento entre conceitos da OWL e da GML.

OWL	GML
Classe	Elemento
Propriedades	Elemento (simples ou complexo) e atributos
Associações	Relacionamentos de hierarquia
Especialização	Derivação de tipos

3 Pré-processamento dos Dados de Entrada

Tanto a ontologia quanto o esquema GML importado devem ser traduzidos em um formato canônico para facilitar a tarefa de determinação das equivalências. Como ambas são estruturas XML, utilizam-se estruturas de dados em árvore como formato canônico de representação. Os *wrappers* indicados na Figura 1 são responsáveis pela geração das árvores canônicas (Figura 2). Um exemplo de atividade de *wrapping* para a ontologia é mostrado a seguir.

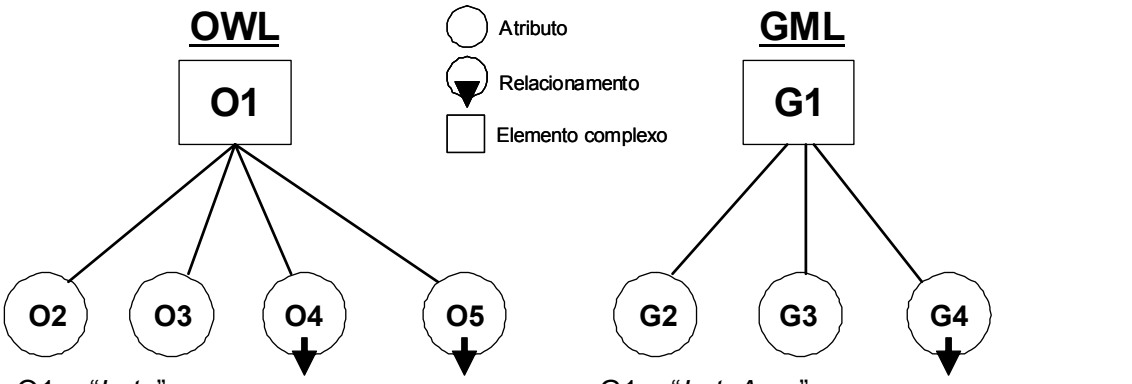
Os dados na ontologia estão organizados em:

(a) definição de classe - considera-se o nome da classe;

(b) propriedades da classe - podem representar atributos simples (*strings*, números etc.) ou complexos (formados por outros atributos) e relacionamentos. Além disso, atributos e relacionamentos podem receber um ou mais valores;

(c) instâncias de objetos - usadas para formar um dicionário de sinônimos.

Na árvore canônica, a descrição da ontologia organiza-se hierarquicamente, ou seja, as propriedades de uma classe C_i (atributos e relacionamentos) são representadas como nodos filhos de C_i. Assim, as classes presentes na ontologia representam nodos na árvore canônica e as propriedades de cada classe tornam-se nodos folha. Além dos atributos, os relacionamentos entre classes também são considerados no cálculo do grau de similaridade.



O1 = "*Lote*"
O2 = "*endereco*" (*string*)
O3 = "*numeroQuarteirao*" (*integer*)
O4 = "*ehParte*" ("*Quarteirao*", atômico)
O5 = "*possuiRepresentacao*" ("*representacaoGeografica*", multivalorado)

G1 = "*LoteArea*"
G2 = "*endereco*" (*string*)
G3 = "*Quadra*" (*integer*)
G4 = "*ehParte*" ("*QuadraCTM*", atômico)

Figura 2. Exemplo de árvore canônica para um elemento OWL e para um elemento GML.

De forma semelhante, uma árvore canônica é gerada para o esquema GML importado. Nesta árvore, cada elemento definido no esquema torna-se um nodo na árvore, com nodos folha representando seus atributos e relacionamentos.

4 Determinação do Grau de Similaridade

Uma vez geradas as representações canônicas, definem-se *graus de similaridade* entre nodos da árvore GML e nodos da árvore da ontologia. Um nodo da árvore GML pode encontrar equivalência com mais de um nodo na árvore da ontologia. Portanto, é necessário estimar um grau de similaridade através de um valor numérico que possa variar entre 0 e 1, de forma a facilitar a determin ação do melhor candidato a similar.

A determinação da similaridade é fortemente baseada em uma abordagem lingüística [Rahm e Bernstein 2001], ou seja, ela determina a equivalência através da igualdade ou similaridade do texto, considera ndo ainda a estrutura dos elementos.

O processamento inicia pela disponibilização de uma lista de sinônimos associados a cada classe da ontologia, no formato [SINÔNIMO, CLASSE, IDIOMA] (Tabela 2). Os sinônimos complementam a ontologia, identificando variações conhecidas na denominação de cada termo. Esta lista de sinônimos é obtida de instâncias da classe *Dicionario* presente na ontologia, mas poderia estar armazenada em um banco de dados. Optou-se por mantê-la na ontologia de forma a centralizar em um único local todo o conhecimento sobre os conceitos no domínio da aplicação.

Tabela 2. Exemplo da lista de sinônimos.

SINÔNIMO	CLASSE	IDIOMA
Lote	Lote	pt
Parcel	Lote	en
Quadra	Quarteirão	pt
Quarteirão	Quarteirão	pt

O identificador de cada nodo representativo de um elemento GML (por exemplo, o nome "Quadra") é primeiramente testado por *igualdade* contra a lista de sinônimos. Caso um ou mais sinônimos correspondentes sejam encontrados (por exemplo, "Lote" = "Lote"), aplica-se uma *métrica de similaridade de estruturas (MCV)* em cada resultado positivo, a fim de se definir o grau de similaridade entre os nodos. Caso contrário, faz-se nova pesquisa na lista de sinônimos, agora aplicando uma *métrica de similaridade entre nomes (MAV)* (por exemplo, "QuadraCTM" = "Quadra"). Caso o grau de similaridade atinja um valor aceitável definido pelo usuário, aplica-se a métrica de similaridade de estruturas.

A opção por fazer primeiro um teste por igualdade contra a lista de sinônimos visa dar mais agilidade ao método, uma vez que este teste é mais rápido do que o uso de métricas.

Diversas métricas de similaridade entre *strings* são encontradas na literatura, como a métrica de *Jaro*, a distância de *Levenshtein* e a distância de *Hamming*. No método proposto, optou-se por usar a métrica *Jaro Winkler* [Chapman 2006]. Esta métrica estende a métrica de *Jaro*, incluindo uma correção do valor final da comparação, de forma a evitar que *strings* diferentes apenas na parte final tenham uma distância grande entre elas. Ela se mostra interessante para o foco deste trabalho, uma vez que considera o conceito de prefixo na comparação das *strings*. Tais situações ocorrem com freqüência em especializações de objetos geográficos, como, por exemplo, *QuadraCTM*, *QuadraProjetada*, *QuadraReal, LoteTributavel* e *LoteCTM*. A Tabela 3 mostra um exemplo da aplicação da métrica *Jaro Winkler*, comparando os resultados obtidos (graus de similaridade) com a distância de *Levenshtein*.

A métrica *Jaro Winkler* é aplicada em métricas MAV mais abrangentes, que consideram também a equivalência entre o tipo de dado de dois atributos sendo comparados. Estas métricas são apresentadas na seqüência.

Tabela 3. Comparação do resultado das métricas *Jaro Winkler* e *Levensthein*.

Ontologia	GML importado	Jaro Winkler	Levenshtein
Quadra	QuadraCTM	0,9556	0,6667
Lote	LoteCTM	0,9143	0,5714
Lote	LoteTributavel	0,8571	0,2857

Pela natureza dos dados manipulados, a determinação da similaridade de estruturas (elementos complexos) necessita adaptar apenas uma *métrica de similaridade de tuplas*. Isto se justifica pelo fato de que a definição de um elemento complexo de um esquema GML é composto pela identificação (nome) do elemento e suas propriedades (atributos e relacionamentos), como uma *tupla* de uma tabela relacional.

A principal métrica para determinar a similaridade de estruturas se resume, neste trabalho, à adaptação da métrica de similaridade de *tuplas* proposta por Dorneles *et al*. (2004):

$$tupleSim(\varepsilon_p, \varepsilon_d) = \frac{\sum_{\varepsilon_p^i.\eta = \varepsilon_d^j.\eta} (sim(\varepsilon_p^i, \varepsilon_d^j))}{\max(m, n)}$$

onde:

- ε_p : um nodo no conjunto *P*;
- ε_d : um nodo no conjunto *D*;
- *n* e *m*: número de filhos de ε_p e ε_d, respectivamente;
- *P*: conjunto de nodos de elementos presentes na árvore do esquema GML;
- *D*: conjunto de nodos de classe presentes na árvore da ontologia.

Através da métrica *tupleSim()* é feita a navegação pela estrutura dos elementos das árvores da ontologia e do esquema GML. Cada nodo ε_p^j filho de ε_p é comparado com um nodo ε_d^i filho de ε_d, com o mesmo nome ($\varepsilon_p^j.\eta = \varepsilon_d^i.\eta$) e o mesmo contexto. A função *max()* retorna o maior número de filhos entre ε_p e ε_d.

A adaptação da métrica, comentada acima, leva em conta que um atributo pode ser simples ou complexo e que relacionamentos são tratados como atributos complexos. No caso de atributo complexo, a métrica é aplicada de uma forma *bottom-up*, ou seja, parte do último nível da árvore canônica do elemento GML, para os níveis superiores. Desta forma, um atributo complexo em um nível superior é tratado como um atributo atômico, pois seu grau de similaridade foi resolvido em uma iter ação prévia.

Além disso, para comparar um nodo filho de um elemento GML com um nodo filho da árvore da ontologia, foi necessário definir neste trabalho duas métricas MAV adicionais, conforme o tipo de nodo encontrado:

- *Métrica para atributos simples*: compara nome dos atributos (métrica *Jaro Winkler*) e seus tipos de dados (análise de compatibilidade):

$$attrSim(\varepsilon_p.\eta, \varepsilon_d.\eta) = \frac{sim(\varepsilon_p.nome, \varepsilon_d.nome) + sim(\varepsilon_p.tipo, \varepsilon_d.tipo)}{2}$$

  onde:

  - ✓ ε_p : nodo filho do elemento da árvore GML;
  - ✓ ε_d : nodo filho do elemento da árvore OWL;
  - ✓ *nome*: identificador do nodo filho;
  - ✓ *tipo*: tipo de dado do nodo filho.

- *Métrica para relacionamentos*: compara nome dos relacionamentos (métrica *Jaro Winkler*) e igualdade de cardinalidade (1:1 – atômico; 1:n - multivalorado):

$$relSim(\varepsilon_p.\eta, \varepsilon_d.\eta) = \frac{sim(\varepsilon_p.nome, \varepsilon_d.nome) + sim(\varepsilon_p.card, \varepsilon_d.card)}{2}$$

onde:

  - ✓ *card*: cardinalidade do nodo filho.

Exemplificando, aplicou-se as métricas *tupleSim()*, *attrSim()* e *relSim()*, aos elementos *Lote* e *LoteArea* da Figura 2. Primeiro, calcularam-se as similaridades entre atributos atômicos:

- *sim2 = attrSim(G2, O2)* = 1
- *sim3 = attrSim(G3, O3)* = 0,78

Posteriormente, calcularam-se as similaridades entre relacionamentos:

- *sim4 = relSim(G4, O4)* = 0,89
- *sim5 = relSim(G5, --)* = 0

Com os resultados anteriores, calculou-se o grau de similaridade final entre os dois elementos:

- *sim1 = tupleSim()* = (1 + 0,78 + 0,89 + 0) / 4 = 0,67

O valor obtido para a métrica *tupleSim()* aplicada ao exemplo, pode ser considerado insuficiente para definir automaticamente os dois elementos como equivalentes. Neste caso, a intervenção do usuário é necessária no processo de avaliação.

5 Catalogação do Mapeamento

Uma vez identificada a similaridade entre nodos classe e nodos elemento nas árvores canônicas dos esquemas OWL e GML, o passo final é armazenar este mapeamento, em duas tabelas de um banco de dados relacional:

- A primeira mantém informações a respeito dos esquemas GML importados (metadados), como identificação do fornecedor do esquema, URL, responsável, versão do esquema, idioma entre outros;
- A segunda mantém as correspondências propriamente ditas. Parte-se do princípio de que para cada elemento do esquema GML principal pode haver um conceito equivalente na ontologia. Assim, são acrescentadas duas novas colunas na tabela para cada esquema GML importado: uma contendo o elemento mapeado e outra contendo o grau de similaridade com o conceito na ontologia. A Tabela 4 apresenta um exemplo desta catalogação.

Tabela 4. Tabela de armazenamento de mapeamentos.

Conceitos Ontologia	GML principal	Similaridade GML principal	GML importado 1	Sim. GML importado 1	...
Quadra	Quadra	1	QuadraCTM	0,96	...
Lote	LoteProjetado	0,86	LoteCTM	0.91	...

6 Considerações Finais

Considerando que a troca de dados geográficos acontece principalmente entre domínios que possuem afinidade entre si, um dado geográfico é mais bem definido semanticamente em um domínio específico do que pela generalização de domínios.

Este artigo propõe uma solução para o problema da interoperabilidade semântica entre esquemas GML no contexto do cadastro urbano. Esta solução prevê a determinação de equivalências de forma semi-automática, usando conhecimento representado como uma ontologia. O método é semi-automático pois não descarta a intervenção do usuário quando um elemento GML pode ser relacionado a vários conceitos da ontologia ou para validar equivalências determinadas automaticamente.

Trabalhos relacionados preocupam-se em traduzir consultas executadas em ambientes heterogêneos fortemente interligados. Este trabalho prevê um outro cenário: municípios de pequeno porte, cujos dados geográficos estão disponíveis em diversas instituições, como prefeituras e companhias de água e esgoto. Neste cenário, considera-se que, isoladamente, estas instituições não teriam condições técnico-financeiras de disponibilizar seus dados. Entretanto, em conjunto, poderiam promover um intercâmbio de dados através de um mecanismo que identifique as semelhanças entre eles. Desta forma, é possível socializar os dados geográficos urbanos e desenvolver novos serviços à comunidade em geral.

Trabalhos futuros prevêem a ampliação do escopo abordado, estendendo o método proposto para resolver a interoperabilidade de esquemas GML para outros domínios; a especificação de um ambiente para consulta a dados remotos, com base nas equivalências; e a integração de instâncias de dados.

Referências Bibliográficas

Brauner, D. F.; Casanova, M. A.; Lucena, C. J. P. Geo-Object Catalogs to enable Geographic Databases Interoperability. In: GEOINFO, 6., 2004, Campos do Jordão. **Anais...** São José dos Campos: INPE, 2004. p. 235-246.

Câmara, G. *et al*. A. Interoperability In Practice: Problems in Semantic C onversion from Current Technology to OpenGIS. In: INTERNATIONAL CONFERENCE ON INTEROPERABLE OPEARTING SYSTEMS, 2., 1999. **Procedings...** Zurich, 1999. p. 120-138.

Chapman, J. **Sam's Strings Metrics**. Disponível em: <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html>. Acessado em: 25 mai. 2006.

Dorneles, C. F. *et al*. Measuring Similarity between Collection of Values. In: ACM INTERNATIONAL WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 6., 2004. **Proceedings...** Washington: ACM, 2004. p. 56-63.

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.

OGC. **Geography Markup Language (GML) Implementation Specification 3.0**. Open GIS Consortium, 2003.

OWL. **Web Ontology Language**. Disponível em: <http://www.w3.org/2004/OWL/>. Acessado em: 16 jul. 2006.

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

Zhang, J. *et al*. Geographic Information Integration and Publishing Based on GML and SVG. In: INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY, 4., 2004. **Proceedings...** Wuhan: IEEE Computer Society, 2004. p. 764-769.

Apêndice – Ontologia de domínio criada no Protégé.