Um *Framework* para Recuperação Semântica de Dados Espaciais

Jaudete Daltio[1,2]**, Carlos Alberto de Carvalho**[3]

[1]Embrapa Gestão Territorial, Campinas – SP – Brasil

[2]Instituto de Computação - Universidade Estadual de Campinas (UNICAMP)
Campinas – SP – Brasil

[3]Escritório de Análise e Monitoramento de Imagens de Satélite do GSI/SAEI/PR
Campinas – SP – Brasil

jaudete.daltio@embrapa.br, calberto@cnpm.embrapa.br

Abstract. *Geographic data represent objects for which the geographic location is an essential feature. Since they represent real-world objects, these data present a lot of intrinsic semantic, which is not always explicitly formalized. Explicit semantic allows higher accuracy in data retrieval and interpretation. The goal of this work is to propose a framework for management and retrieval of geographic data, combining semantic and spatial aspects. The main contributions of this work are the specification and implementation of the proposed framework.*

Resumo. *Dados geográficos representam objetos para os quais a localização geográfica é uma característica essencial para sua análise. Por representarem objetos do mundo real, esses dados possuem muita semântica intrínseca, que nem sempre é explicitamente formalizada. A semântica explícita possibilita maior acurácia na recuperação e interpretação dos dados. O objetivo deste trabalho é propor um framework para recuperação de dados geográficos que manipule aspectos semânticos e espaciais de forma integrada. Dentre as contribuições estão a especificação e a implementação do framework proposto.*

1. Introdução e Motivação

Sistemas de Informações Geográficas (SIGs) são sistemas capazes de manipular dados georreferenciados, ou seja, dados que representam fatos, objetos e fenômenos associados à uma localização sobre a superfície terrestre. Para estes objetos, a localização geográfica é uma característica inerente à informação e indispensável para analisá-la [Câmara et al. 1996]. Além de dados alfanuméricos, esses sistemas correlacionam dados espaciais vetoriais e matriciais.

Por representarem objetos do mundo real, dados geográficos possuem muita semântica intrínseca, nem sempre explicitamente formalizada. A interpretação dos dados é, em geral, responsabilidade dos especialistas do domínio. Em grupos de trabalho dispersos, esses especialistas podem possuir metodologias, focos de pesquisa e vocabulários distintos. Esse problema ganha maior dimensão para imagens de satélite, que possuem muitas informações agregadas e demandam elevado processamento computacional para sua interpretação, como classificação e reconhecimentos de padrões.

Ontologias vêm se materializando como a principal tecnologia para representação de semântica [Horrocks 2008]. Tratam-se de estruturas computacionais capazes de representar os termos de um domínio e seus relacionamentos. Seu uso tem sido cada vez mais difundido em geotecnologias, modelando desde atributos e metadados à relacionamentos espaciais. A associação de semântica ainda representa um dos três principais desafios a serem superados pela nova geração de SIGs [Câmara et al. 2009].

O objetivo deste trabalho é especificar e implementar um *framework* para gerenciamento de dados geográficos, integrando aspectos semânticos e espaciais. O *framework* será capaz de propagar a semântica entre os dados geográficos – de vetoriais para matriciais – considerando suas correlações espaciais. Com isso, será possível incorporar aspectos semânticos às imagens de satélite e auxiliar seu processo de recuperação. Serão utilizadas ontologias como base para as anotações semânticas.

O restante desse artigo segue a seguinte organização: a seção 2 apresenta os aspectos de pesquisa relacionados ao trabalho. A seção 3 descreve o *framework* proposto, seus aspectos de implementação e estudos de caso que validam a aplicabilidade da solução proposta. A seção 4 apresenta os resultados e as contribuições previstas para o trabalho.

2. Aspectos de Pesquisa Envolvidos

Os aspectos de pesquisa desse trabalho são: anotações, semântica (ontologias) e ferramentas de anotação semântica. As seções subsequentes detalham esses tópicos.

2.1. Fundamentação Teórica - Anotações e Semântica

Anotar é o processo de adicionar notas ou comentários a um dado. De forma análoga aos metadados, uma anotação é utilizada para descrever um dado, ou parte dele, adotando ou não um vocabulário de referência. O termo "anotação semântica" decorre do uso de ontologias como vocabulário de referência para a anotação [Macário 2009], visando interoperabilidade. Em aplicações geográficas, uma anotação também pode considerar o componente espacial. O diferencial das anotações semânticas está no processo de recuperação. Mecanismos tradicionais de busca por palavras-chave possuem muitas limitações e a análise do contexto pode melhorar a acurácia deste processo.

Ontologias são especificações explícitas de uma conceitualização – uma definição consensual a respeito da representação de um domínio. O domínio geográfico possui várias ontologias e, considerando os dados utilizados neste trabalho, selecionou-se ontologias adequadas para a representação de empreendimentos de infraestrutura governamental e dos recursos naturais a cerca deles. São elas:
- **AGROVOC** [1]: descreve a semântica de temas como agricultura, silvicultura, pesca e outros domínios relacionados com alimentação, como meio ambiente;
- **SWEET** [2]: termos sobre dados científicos, com conceitos ortogonais como espaço, tempo, quantidades físicas, e de conhecimento científico, como fenômenos e eventos;
- **VCGE** [3]: padrão de interoperabilidade para facilitar a indexação de conteúdo nos portais governamentais, tratando de assuntos de interesse do setor público;
- **OnLocus [Borges 2006]**: conceitos no domínio espaço geográfico urbano, feições naturais, objetos e lugares, incluindo os relacionamentos entre eles.

[1] http://aims.fao.org/standards/agrovoc
[2] http://sweet.jpl.nasa.gov/ontology/
[3] http://vocab.e.gov.br/2011/03/vcge

2.2. Trabalhos Correlatos - Ferramentas de Anotação

A Figura 1 apresenta algumas ferramentas citadas na literatura para a anotação semântica: KIM [Popov et al. 2003], E-Culture [Hollink 2006], CREAM [Handschuh and Staab 2002], OnLocus [Borges 2006] e Macario [Macário 2009] e o *framework* proposto. Como mostra a figura, três dessas ferramentas consideraram aspectos espaciais, diversas fontes de dados web e utilizam o formato RDF/OWL para representar as anotações. No *framework* proposto, as anotações são armazenadas em BD relacionais utilizando conceitos de ontologias OWL, o processo de anotação é manual para os dados vetoriais e a propagação dessas anotações é automática. O diferencial da proposta está nesse processo de propagação, considerando correlações espaciais, e no processo de recuperação dos dados buscando por relacionamentos entre os vocabulários utilizados na consulta e nas anotações.

Ferramenta	**KIM**	**E-Culture**	**CREAM**	**OnLocus**	**Macario**	**Proposta**
Formato	RDF/OWL	RDF/OWL	RDF/OWL	XML	Tripla <s,m,o>	Tuplas com OWL
Processamento	Automático	Semi-automático	Automático	Automático	Semi-automático	Manual e Automático
Dados Origem	Páginas Web	Imagens	Páginas Web, vídeos e imagens	Páginas Web	Dados geográficos Web	Dados geográficos vetoriais e raster
Espacial	Não	Sim	Não	Sim	Sim	Sim

Figura 1. Comparativo entre Ferramentas de Anotação

3. Trabalho Proposto

O objetivo do *framework* proposto é prover a recuperação semântica de dados geográficos. Essa recuperação será viabilizada pela construção de anotações semânticas, pela propagação dessas anotações entre os objetos geográficos (vetoriais e matriciais) e por mecanismos de consulta que permitam correlacionar essas anotações. O *framework* utiliza ontologias do contexto geográfico como base para a elaboração de anotações semânticas. Essas ontologias são manipuladas pelo Aondê, um serviço de ontologias capaz de prover acesso, manipulação, análise e integração de ontologias [Daltio and Medeiros 2008]. O Aondê é composto por duas principais camadas, encapsuladas em serviços Web: *Repositórios Semânticos*, responsável pelo gerenciamento das ontologias e seus metadados, e *Operações*, responsável pelas funcionalidades como busca e *ranking*, consultas e integração de ontologias.

A Figura 2 ilustra a arquitetura do *framework*, composto por duas camadas: **Repositórios de Dados** e **Camada de Recuperação**. As funcionalidades são acessadas via **Interface Web**. O **Repositórios de Dados** possui por dois catálogos dedicados ao armazenamento de dados geográficos e um para o armazenamento das anotações semânticas. A **Camada de Recuperação** provê a inclusão de dados nos repositórios, a propagação das anotações entre os dados geográficos e mecanismos de consulta. A figura mostra ainda que as interações entre o *framework* e o serviço de ontologias Aondê ocorrem nessa camada. Os parágrafos subsequentes descrevem essas camadas.

Repositório de Dados: responsável pela persistência dos dados. Os dados matriciais (imagens de satélite) são armazenados via sistema de arquivos, seus metadados e

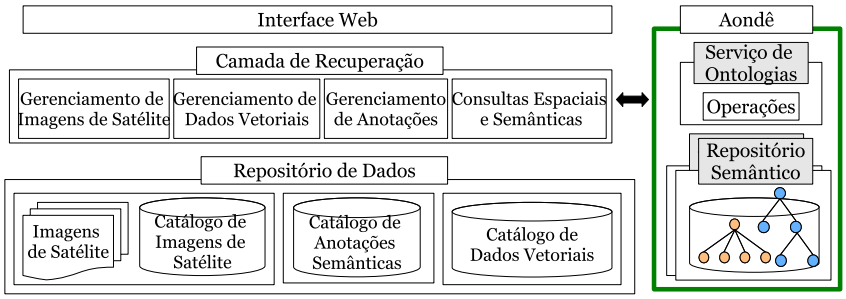

Figura 2. Arquitetura do *Framework* Proposto

retângulos envolventes no catálogo de imagens. As cenas de regiões contínuas, de mesma data e sensor, são agrupadas em mosaicos no catálogo. O catálogo de dados vetoriais armazena a geometria de empreendimentos governamentais de infraestrutura (aeroportos, usinas hidrelétricas, dentre outros), acrescidos de metadados (como divisão territorial). O catálogo de anotações semânticas armazena as anotações, materializando o link entre os dados espaciais e conceitos de ontologias (triplas RDF/OWL).

Camada de Recuperação: composta pelos módulos:

Gerenciamento de Imagens de Satélite: provê a inclusão de imagens de satélite no *framework*, criando registros no catálogo de imagens associados aos arquivos de imagens.

Gerenciamento de Dados Vetoriais: provê a inclusão de empreendimentos, pela inserção dos dados vetoriais, textuais e cruzamento com dados espaciais complementares.

Gerenciamento de Anotações: provê a criação e propagação de anotações. O processo de anotação possui duas entradas: o empreendimento e o termo de interesse. A partir desse termo, o *framework* utiliza o Aondê (operação busca e *rank*) para a seleção da ontologia. Essa operação foi estendida para retornar o conceito mais representativo deste termo na ontologia. Com isso, cria-se uma anotação associando o empreendimento em questão a essa tripla RDF/OWL (e sua ontologia de origem). A propagação da anotação é feita criando-se novas associações entre esse conceito da ontologia com as imagens de satélite cujos retângulos envolventes possuam interseção espacial com esse empreendimento.

Consultas Espaciais e Semânticas: provê mecanismos para recuperação combinando aspectos espaciais e semânticos. Há três opções de entrada: um empreendimento, uma imagem de satélite e um termo de interesse. Para os dois primeiros, são disponibilizados os metadados para filtragem e, ao retornar-se um resultado que atenda ao padrão de consulta, utiliza-se interseção espacial para retornar outros dados espacialmente relacionados. Para o terceiro caso, utiliza-se o Aondê para encontrar conceitos em ontologias que representem ocorrências do termo de consulta e esse resultado é comparado com o catálogo de anotações. A estratégia de recuperação possui três níveis de busca: **(i) busca direta:** retorna os registros de dados anotados com algum dos resultados retornados, sendo possível combinar termos diferentes na busca pelas anotações;**(ii) busca indireta:** retorna os registros de dados anotados com alguma das ontologias retornadas no resultado, ordenando-se o resultado pela distância entre os termos (consulta e anotação); **(iii) busca por alinhamento:** utiliza-se o Aondê para alinhar cada par de ontologias (ontologia que contém o termo buscado + ontologia usada na anotação). Caso algum

alinhamento seja encontrado, o procedimento de recuperação de dados é análogo à busca indireta e as ontologias alinhadas são manipuladas como se fossem uma ontologia única.

Interface Web: camada de visualização do *framework*. Foram desenvolvidas interfaces para a visualização de empreendimentos e imagens de satélite (e suas correlações espaciais). A interface para manipulação das anotações semânticas está em fase de elaboração e propõe-se a adoção de árvore hiperbólica para visualização.

3.1. Aspectos de Implementação

O protótipo do *framework* está em fase de implementação. O **Repositório de Dados** utiliza o SGBD PostgreSQL e a extensão PostGIS [4] para manipulação dos dados geográficos. Foram desenvolvidos *scripts* para a inserção automática de imagens e empreendimentos. A **Camada de Recuperação** e a **Interface Web** estão sendo implementadas em PHP. Para a publicação e navegação nos dados espaciais utiliza-se o servidor de mapas MapServer [5] e o servidor Web Apache. A manipulação das ontologias é responsabilidade do Aondê, acessado via serviços Web.

3.2. Estudo de Caso

Para esse estudo de caso, utilizou-se um conjunto de empreendimentos governamentais de infraestrutura imageados entre 2005 e 2012. As ontologias descritas na seção 2.1 foram aplicadas como vocabulário de anotação. Para a usina hidrelétrica Estreito (polígono), localizada no Rio Tocantins, pesquisou-se os termos para anotação: *hidrelétrica, barragem, rio, energia, Maranhão, Tocantins*, e anotações foram criadas a partir dos resultados: *geracao-energia-hidreletrica* (VCGE), Dam (SWEET), *Energia hidroeléctrica, Rio, Maranhão* (AgroVOC). Para a rodovia BR-153 (linha), que atravessa o estado de Tocantins, pesquisou-se os termos: *rodovia, estrada e transporte*, e anotações foram criadas a partir dos resultados: *Infraestrutura de transporte rodoviário* (VCGE), *rodovia, construção de estradas e Transporte rodoviário* (AgroVOC). Todas as anotações propagadas para as imagens de satélite com interseção espacial nesses empreendimentos.

Considere a consulta ao *framework*: *"Retorne imagens de satélite de rios a partir de 2008"*. O termo de consulta *rio*, ao ser buscado no Aondê, irá retornar um dos conceitos utilizados na anotação da hidrelétrica (AgroVOC), logo todas as imagens para as quais essa anotação foi propagada serão retornadas por busca direta. Essas imagens serão filtradas pelo metadado "data de imageamento", retornando apenas as que atendem ao critério de data superior à 2008. O mesmo ocorreria com qualquer termo de consulta utilizando algum dos termos de anotação. Uma consulta mais elaborada poderia envolver dois ou mais conceitos de anotação: *''Retorne imagens de satélite de rodovias e rios"*. Neste caso, o mesmo processo de busca também seria feito com o termo *rodovias* e seriam retornadas as imagens que possuíssem ambas anotações.

Considere uma consulta mais geral: *"Retorne imagens de satélite de água"*. O termo de consulta *água* irá retornar o conceito *águas*, dois níveis acima do conceito *geracao-energia-hidreletrica* na VGCE e, com isso, as imagens com as anotações da hidrelétrica serão retornadas por busca indireta. Essas imagens serão penalizadas no *ranking* por essa distância de 2 termos. Outro possível caminho de indexação ocorre

[4] http://postgis.refractions.net/
[5] http://mapserver.org/

pelo conceito *BodyOfWater* (um nível acima do conceito *Dam*). Caso imagens diferentes estivessem anotadas com esses conceitos, as anotadas com *Dam* seriam mostradas primeiramente. Um outro exemplo de consulta seria: *"Retorne as imagens de satélite de avenidas"*. O termo *avenida* irá retornar a ontologia OnLocus, que não foi utilizada em nenhuma anotação. Porém, o Aondê é capaz de alinhar essa ontologia com a AgroVOC, retornando, por busca por alinhamento, as imagens anotadas com o termo *rodovia*.

4. Resultados Esperados e Contribuições

Este trabalho atende uma demanda recorrente no gerenciamento de dados geográficos: a explícita associação de semântica aos dados e a incorporação dessa característica em mecanismos de consulta. A assertividade na recuperação dos dados terá influência direta de dois principais fatores: a precisão da anotação criada e a especificidade da ontologia utilizada nessa anotação. Quanto mais ricas e específicas forem as ontologias de origem, maiores serão as possibilidades de exploração dos relacionamentos entre os termos no domínio de interesse e de alinhamentos com outras ontologias complementares.

As principais contribuições esperadas deste trabalho são: (i) levantamento de ontologias utilizadas na representação de dados geográficos; (ii) análise das estratégias de anotação semântica e (iii) especificação e implementação de um *framework* para anotação e recuperação semântica de dados espaciais. A continuidade do projeto prevê a inclusão da dimensão temporal na geometria dos dados vetoriais e a exploração de outros relacionamentos espaciais, além da sobreposição. Além disso, prevê-se a adoção dos padrões de metadados reconhecidos para infraestruturas de dados espaciais [6].

Referências

Borges, K. A. V. (2006). *Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Web*. PhD thesis, UFMG.

Câmara, G., Casanova, M. A., Hemerly, A. S., Magalhães, G. C., and Medeiros, C. M. B. (1996). *Anatomia de sistemas de informações geográficas*. INPE, S. J. dos Campos.

Câmara, G., Vinhas, L., Davis, C., Fonseca, F., and Carneiro, T. G. S. (2009). Geographical information engineering in the 21st century. In *Research Trends in GIS*, pages 203–218. Springer-Verlag, Berlin Heidelberg.

Daltio, J. and Medeiros, C. B. (2008). Aondê: An ontology web service for interoperability across biodiversity applications. *Inf. Syst.*, 33(7-8):724–753.

Handschuh, S. and Staab, S. (2002). Authoring and annotation of web pages in cream. In *WWW '02: Proc. 11th Int. Conf. WWW*, pages 462–473, NY, USA. ACM.

Hollink, L. (2006). *Semantic Annotation for Retrieval of Visual Resources*. PhD thesis, Vrije Universiteit Amsterdam.

Horrocks, I. (2008). Ontologies and the semantic web. *Commun. ACM*, 51(12):58–67.

Macário, C. G. N. (2009). *Semantic Annotation of Geospatial Data*. PhD thesis, IC - Unicamp.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., and Goranov, M. (2003). Kim - semantic annotation platform. In *ISWC 2003*, pages 834–849. Springer Berlin.

[6] http://www.inde.gov.br/