1
Introdução

1.1.
Motivação

Dados geográficos estão disponíveis em uma grande variedade de repositórios, desde os computadores pessoais até repositórios sofisticados mantidos por organizações. Para ajudar na localização e acesso aos dados disponíveis, um tipo de solução frequentemente utilizado é organizar descrições dos dados em catálogos de metadados.

Apesar da óbvia utilidade dos catálogos de metadados, o processo de gerá-los manualmente pode ser tedioso ou até impossível, dependendo da quantidade de metadados envolvida. Por isso, catálogos devem possuir um componente que automatiza o processo de geração e armazenamento de metadados tanto quanto possível. Tal componente pode obter metadados tanto por uma análise do recipiente onde se encontra armazenado o dado a ser catalogado, como dos dados já armazenados no próprio catálogo ou em outro repositório disponível.

Em se tratando do domínio de sistemas de informação geográfica, temos duas valiosas ferramentas para abordar esse problema. Primeiro, vários sistemas de geo-referenciamento associam a cada entidade geográfica sua posição na superfície terrestre, o que funciona como um identificador universal para a entidade, ou pelo menos uma aproximação. Em segundo, muitos dicionários de nomes geográficos têm sido desenvolvidos e disponibilizados na *Web* nos últimos anos.

É possível generalizar a solução do problema de localização e acesso a dados geográficos para outros tipos de dados. A diferença encontra-se nas técnicas utilizadas para a geração dos metadados.

Neste trabalho propomos uma estratégia de geração e catalogação automática de metadados e uma arquitetura de software para implementá-la. A arquitetura foi concebida de maneira que gerasse como subproduto um *framework* para catalogação automática de dados de quaisquer naturezas. Não abordamos, no entanto, estratégias de geração de metadados para dados não geográficos.

1.2. Trabalhos relacionados

Para encaminhar o problema da catalogação de dados geográficos, inúmeras soluções têm sido propostas na comunidade científica. A seguir são brevemente apresentados alguns trabalhos relacionados que serviram e podem vir a servir, sobremaneira, como inspiração e fontes de informação para o desenvolvimento desta pesquisa.

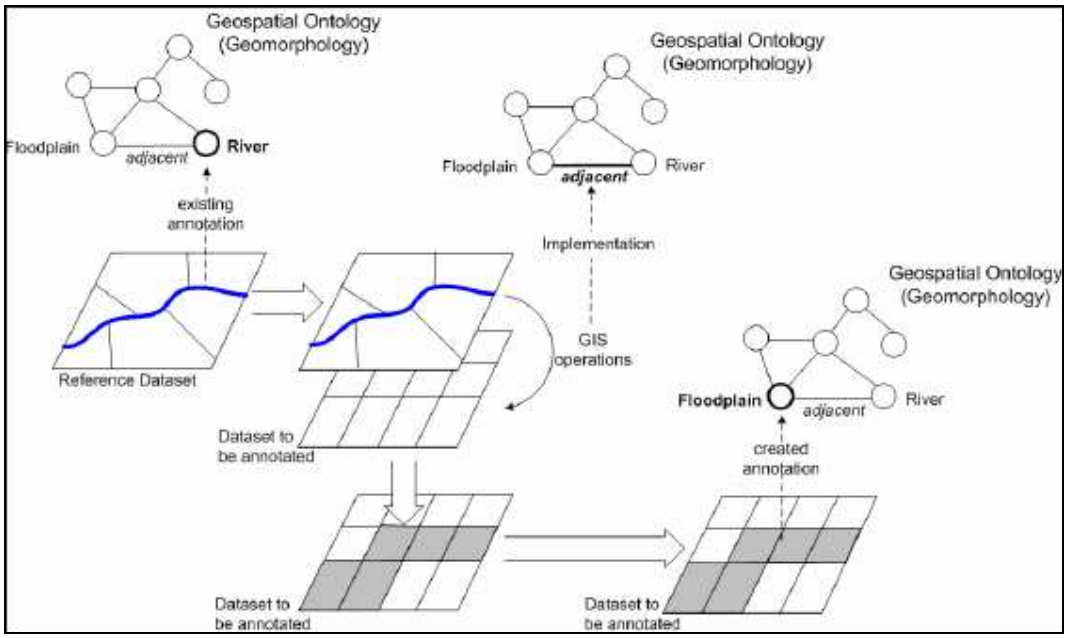

Figura 1 – Esquema do processo de identificação de áreas de alagamento ao redor de um rio

Em *The Role of Spatial Relations in Automating the Semantic Annotation of Geodata*, Klien e Lutz [2] propõem um procedimento para descrição semi-automática de dados geográficos com o propósito de identificar áreas de alagamento ao redor de rios. As descrições são produzidas a partir da identificação de relacionamentos topológicos, através de operações GIS, entre dados de referência e dados que se deseja descrever. Em seguida, uma

ontologia geo-espacial é utilizada para classificar os dados e associá-los entre si através dos relacionamentos topológicos identificados. Por fim, técnicas de mineração de dados sobre a ontologia e seus dados são empregadas para identificar as áreas de interesse. Um esquema dessa abordagem pode ser visto na Figura 1.

Em *Semantic Annotation of Image Collections*, Hollink et al. [4] apresentam uma ferramenta para registro de descrições e busca de dados em uma coleção de imagens de arte. As descrições são produzidas baseadas em um vocabulário restrito e definido por quatro dicionários: *The Art and Architecture Thesaurus (AAT)*, *WordNet*, *IconClass* e *The Union list of Artist Names (ULAN)*. O esquema de metadados é derivado do *VRA 3.0*, que é uma especialização do *Dublin Core* para aplicações de imagens de arte.

A qualidade e precisão das descrições baseiam-se em duas técnicas. Na primeira, o domínio de cada atributo da imagem é limitado a um vocabulário especifico de um ou mais dos quatro dicionários, dessa forma padroniza-se o vocabulário empregado. Por exemplo, o nome do artista autor da obra é selecionado, pela ferramenta, do *ULAN*, a descrição da cena pode ser obtida no *IconClass*, etc. A segunda técnica presta-se a registrar sentenças de descrição da obra. É proposta uma sintaxe na forma: “agente – ação – objeto – recipiente”. Um exemplo dessa abordagem é um quadro de Chagal que poderia ser descrito pela seguinte sentença: “*Chagal, Mark kiss wives”*. Também aqui, é empregado um vocabulário controlado oferecido pelos dicionários.

Para o mecanismo de busca é proposto o alinhamento dos dicionários pela implementação de relacionamentos entre termos. Três tipos de relacionamentos são definidos: equivalência, subclasse e domínio específico. Uma equivalência ocorre quando dois termos de dicionários diferentes são equivalentes. Uma subclasse é quando um termo em um dicionário define um conceito que pode ser visto como uma subclassificação de um conceito em outro dicionário. Por fim, o domínio específico mapeia conceitos que, embora diferentes, estão relacionados de alguma forma, por exemplo, técnicas de pintura e materiais utilizados. Essa técnica permite utilizar o vocabulário de todos os dicionários para formular-se consultas aos dados.

Em *The Role of Gazetteers in Geographic Knowledge Discovery on the Web*, Ligiane Souza et al. [3] descrevem a arquitetura do dicionário geográfico *LOCUS* e seu mecanismo de consulta. Propõem que essa ferramenta seja utilizada em conjunto com motores de busca de páginas *Web* para permitir a seleção de dados com base em informações geográficas. Apresentam, também, uma estratégia para popular o dicionário a partir de páginas *Web* disponíveis. Um agente percorre várias URL's procurando por informações que se encaixam em padrões previamente especificados pelo usuário e extraem deles fragmentos de informação para compor uma entrada do dicionário. Exemplificam o processo com a *home page* de um hotel da qual são extraídos o nome do hotel, seu endereço e telefones. O endereço capturado é, então, cruzado com informações prévias contidas no dicionário para agregar-lhe referências geográficas e finalmente cadastrado no dicionário.

Em *GeoReferencing the Semantic Web: ontology based markup of geographically referenced information*, Hiramatsu et al. [5] apresentam duas ferramentas para tratamento de informações geo-referenciadas. A primeira permite ao usuário editar graficamente formas geométricas geo-referenciadas e produz um arquivo RDF contendo todos os relacionamentos topológicos entre elas. Para esse processo, Hiramatsu desenvolveu uma ontologia de descrição de relacionamentos topológicos e outra para classificação de locais geográficos, como estado, continente, país, etc. A segunda ferramenta seleciona, em um repositório de objetos, duas entradas e calcula o relacionamento topológico entre elas. Os objetos são previamente descritos em RDF, utilizando a ferramenta anterior, e armazenados no repositório. Serviços *Web* para consulta ao repositório e cálculo de relacionamento foram especialmente desenvolvidos.

1.3. Organização do trabalho

Esse trabalho está organizado da seguinte forma. O capítulo 2 apresenta as principais tecnologias utilizadas para catalogação automática de dados geográficos: catálogos de metadados e dicionários geográficos. Para catálogos, são indicados padrões internacionais de esquemas de dados e interfaces de serviços. Para dicionários, são descritos brevemente alguns dicionários existentes e feita uma análise mais detalhada para o *ADL Gazetteer*. O capítulo 3 especifica um método para correlacionar dados geográficos com objetos de um

dicionário de modo a obter uma descrição do dado. Apresenta, também, uma arquitetura de software para aplicar esse método na catalogação automática dos dados. O capítulo 4, como prova dos conceitos apresentados, especifica um projeto de *software* para catalogação automática de dados geográficos. Ao final do capítulo os conceitos de catalogação automática são generalizados em um *framework* que pode ser customizado para tratar outros tipos de dados. Por fim, o capítulo 5 apresenta algumas conclusões sobre esta dissertação e indica algumas linhas de pesquisa a serem desenvolvidas a partir do exposto aqui.