Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

TESE DE DOUTORADO

SESDI: Um Arcabouço para a Recuperação de Dados Geográficos em Infraestruturas de Dados Espaciais

Fabio Gomes de Andrade
(Doutorando)

Cláudio de Souza Baptista, Ph. D.
(Orientador)

CAMPINA GRANDE – PB
2012

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

SESDI: Um Arcabouço para a Recuperação de Dados Geográficos em Infraestruturas de Dados Espaciais

Fabio Gomes de Andrade

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação

Linha de Pesquisa: Sistemas de Informação e Bancos de Dados

Cláudio de Souza Baptista, Ph. D.

(Orientador)

Campina Grande, Paraíba, Brasil

A553s Andrade, Fabio Gomes de
SESDI : um arcabouco para a recuperacao de dados geograficos em infraestruturas de dados espaciais / Fabio Gomes de Andrade. - Campina Grande, 2012.
145 f. : il.

Tese (Doutorado em Ciencia da Computacao) - Universidade Federal de Campina Grande, Centro de Engenharia Eletrica e Informatica.

1. Infraestruturas de Dados Espaciais 2. Recuperacao da Informacao Geografica 3. Ranking 4. Web Semantica 5. Ontologias 6. Tese I. Baptista, Claudio de Souza, Dr. II. Universidade Federal de Campina Grande - Campina Grande (PB) III. Título

CDU 004.6(043)

"SESDI: UM ARCABOUÇO PARA A RECUPERAÇÃO DE DADOS GEOGRÁFICOS EM INFRAESTRUTURAS DE DADOS ESPACIAIS"

FABIO GOMES DE ANDRADE

TESE APROVADA EM 27/08/2012

CLÁUDIO DE SOUZA BAPTISTA, Ph.D
Orientador(a)

ULRICH SCHIEL, Dr.
Examinador(a)

CARLOS EDUARDO SANTOS PIRES, Dr.
Examinador(a)

VALERIA CESARIO TIMES, Ph.D
Examinador(a)

CLODOVEU AUGUSTO DAVIS JUNIOR, Dr.
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Nos últimos anos, infraestruturas de dados espaciais (IDE) têm conquistado uma grande popularidade como uma solução para facilitar a interoperabilidade e o acesso a dados geográficos oferecidos por diferentes organizações. Entretanto, os serviços de catálogo oferecidos atualmente por estas infraestruturas possuem algumas limitações que tornam difícil para o usuário localizar os dados geográficos que já são oferecidos pela IDE. Algumas das limitações dos catálogos atuais incluem o uso de um único registro de metadados para descrever todas as camadas oferecidas por um serviço, o uso de palavras-chave para a resolução de consultas temáticas e a falta de ranking para organizar os resultados obtidos a partir de uma consulta. Visando superar estas limitações, esta tese propõe SESDI (Semantically Enabled Spatial Data Infrastrucutres), um arcabouço que usa ontologias e técnicas da recuperação da informação clássica para melhorar a localização de dados geográficos oferecidos por uma IDE. Ademais, o arcabouço propõe medidas de ranking para a resolução de consultas espaciais, temáticas, temporais e globais.

Abstract

In recent years, spatial data infrastructures (SDIs) have gained great popularity as a solution to facilitate the interoperability and the access to geographic data offered by different agencies. Nevertheless, their current catalogs have several limitations that make difficult for user to find the geographic data that are currently offered by the SDI. Some current catalog drawbacks include the use of just one record to describe all the *feature types* offered by a service, the use of keywords to solve semantic queries and the lack of a ranking metric to organize the results retrieved from a query. Aiming to overcome these limitations, this thesis proposes SESDI (Semantically Enabled Spatial Data Infrastrucutres), which is a framework that uses ontologies and techniques of classic information retrieval to improve geographic data retrieval in a SDI. Moreover, the framework proposes several ranking metrics to solve spatial, semantic, temporal and global queries.

Agradecimentos

Primeiramente a Deus, fonte maior de amor e misericórdia.

À CAPES, pelo subsídio financeiro.

À minha esposa Ranielly, por todo o apoio e paciência ao longo destes quatro anos e à minha filha Isabella, o anjo da guarda que me faz esquecer de todos os meus problemas.

Aos meus pais José e Alzira, por todo o amor e carinho que me dedicaram por toda a minha vida, e pela companhia durante as viagens para Campina Grande.

Às minhas irmãs Sheila e Fernanda, pelo incentivo, aos meus cunhados Gean e Camilo, e aos meus sobrinhos Gabriella, Lívia, Camille, Penélope e Alexandre.

À minha sogra Auzenir e ao meu tio Zezinho, pelo grande incentivo.

Ao meu orientador Cláudio Baptista, pela orientação criteriosa, pela paciência, pelo apoio e pelos conhecimentos transmitidos.

Aos meus colegas de laboratório Fábio Leite, Hugo, Damião, Yuri, Tiago e Ana Gabrielle, por todo o apoio e pelos ótimos momentos no LSI.

A todos os meus colegas de trabalho, especialmente Claudivan, Carlos, Daladier, André, Ricardo, Nadja, Valéria, Gastão e Valnyr, que sempre me apoiaram e contribuíram para o meu afastamento.

À COPIN, pela oportunidade de cursar o doutorado.

A todas as pessoas que, de alguma forma, contribuíram para a realização deste trabalho.

"Não é mérito o fato de nunca termos caído, mas o de termos nos levantado todas as vezes em que caímos"

Provérbio Árabe

Conteúdo

Lista de Símbolos

ASDI	Australian Spatial Data Infrastructure
CONCAR	Comissão Nacional de Cartografia
CEMG	Comitê de Estruturação de Metadados Geoespaciais
CS-DGM	Content Standard for Digital Geospatial Metadata
CSW	Catalog Service for the Web
DAML	DARPA Agent Markup Language
DAML-S	DARPA Agent Markup Language for Services
DC	Dublin Core
DTD	Document Type Definition
ESMI	European Spatial Metadata Infrastructure
EUROGI	European Umbrella Organisation for Geographic Information
FGDC	Federal Geographic Data Committee
GML	Geography Markup Language
GSDI	Global Spatial Data Infrastructure
HTML	HyperText Markup Language
INDE	Infraestrutura Nacional de Dados Espaciais
IR	Information Retrieval
ISO	International Organization for Standardization
MGB	Metadados Geoespaciais do Brasil
NASA	National Aeronautics and Space Administration
NSDI	National Spatial Data Infrastructure
OGC	Open Geospatial Consortium
OIL	Ontology Inference Layer
OWL	Web Ontology Language
OWL-S	Web Ontology Language for Services
RDF	Resource Description Framework
REST	Representational State Transfer
RQL	RDF Query Language
SOAP	Simple Object Access Protocol
SGBD	Sistema de Gerência de Bancos de Dados
SDI	Spatial Data Infratructures

SVG	Scalable Vector Graphics
SWEET	Semantic Web for Earth and Environmental Terminology
SWRL	Semantic Web Rule Language
UDDI	Universal Description, Discovery and Integration
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	Extensible Markup Language
XSLT	eXtensible Stylesheet Language for Transformation
WCS	Web Coverage Service
WFS	Web Feature Service
WMS	Web Map Service
WPS	Web Processing Service
WSDL	Web Services Description Language
WSMO	Web Services Modeling Ontology
W3C	World Wide Web Consortium

Lista de Figuras

Lista de Tabelas

Capítulo 1 - Introdução

1.1 Motivação

Nos últimos anos, a *web* tem se tornado um grande repositório de dados geográficos. Este tipo de informação possui uma grande importância no processo de tomada de decisões em várias áreas, como o planejamento urbano, a gestão de recursos naturais e o gerenciamento de desastres. Atualmente, dados geográficos são disponibilizados por diversos tipos de provedores, tais como agências públicas de diversos níveis de governo, empresas privadas, instituições acadêmicas e pessoas comuns. Da mesma forma, as informações disponibilizadas podem ser acessadas e utilizadas por vários tipos de clientes, que variam desde agentes de organizações de diversos níveis de governo a usuários casuais. Apesar da grande quantidade de informação disponível, a localização e utilização destes dados ainda representam tarefas difíceis de serem realizadas.

Uma questão importante que dificulta a realização destas tarefas é que os dados são oferecidos de forma heterogênea. Tal problema acontece porque os mesmos são disponibilizados por fontes autônomas, que atuam em contextos distintos, possuem necessidades e perspectivas diferentes e usam diferentes formatos para a representação e documentação dos seus dados. O principal problema causado por esta heterogeneidade consiste na dificuldade em integrar dados disponibilizados por diferentes fontes de informação. Tal limitação, aliada à dificuldade em localizar os dados que já estão disponíveis, faz com que muitas agências ainda tenham que gastar muito tempo e dinheiro na produção de dados geográficos que já são disponibilizados por outras organizações, e que poderiam ser reutilizados sem custos ou a custos bem mais baixos.

O problema da heterogeneidade fez surgir a necessidade de se desenvolver uma série de padrões para a área de dados geográficos, de forma a aumentar a interoperabilidade entre os dados e aplicações existentes. Esta necessidade levou à criação do OGC – *Open Geographic Consortium*, que é um consórcio formado por mais de quatrocentas organizações de diversos tipos, como universidades, empresas privadas e organizações não governamentais. Desde a sua proposição, o OGC vem desenvolvendo uma série de padrões para o domínio geoespacial. Dentre estes padrões, pode-se destacar uma série de serviços para o acesso a dados geográficos. Estes serviços permitem acessar, de forma padronizada, dados geográficos oferecidos em

diferentes formatos. Exemplos desses serviços incluem o *Web Map Service* (WMS) (OGC, 2004), que padroniza o acesso a camadas de mapas vetoriais, e o *Web Feature Service* (WFS) (OGC, 2005a), que permite a recuperação de feições espaciais em diferentes formatos, principalmente o GML, que é uma linguagem baseada em XML para a descrição de dados geográficos.

Uma característica importante dos serviços de acesso a dados geográficos é que cada serviço oferece acesso a um conjunto de dados oferecidos por um provedor. Esses dados são oferecidos na forma de camadas, que são chamadas de *feature types*. Cada camada fornece informação sobre um determinado tema, como clima, hidrografia, relevo, estradas, entre outros. Por exemplo, um serviço WMS que oferece dados sobre a hidrografia do Brasil pode oferecer camadas sobre corpos hídricos, rios, bacias hidrográficas, reservatórios, entre outras.

Por sua vez, cada camada oferece um conjunto de dados georeferenciados sobre o seu respectivo tema. Estes dados são chamados de feições (do inglês, *features*). Por exemplo, um *feature type* que oferece dados sobre os rios da hidrografia do Brasil pode oferecer feições referentes ao Rio Amazonas, Rio São Francisco, Rio Solimões, entre outros. Cada feição é composta por uma geometria, que descreve a sua localização, e um conjunto de atributos, que descrevem as suas propriedades. Por exemplo, cada rio, além de sua geometria, pode ter atributos que descrevem o seu nome, a sua extensão, a sua área e o seu volume de água.

Os serviços propostos pelo OGC facilitaram o acesso a dados geográficos oferecidos por diferentes organizações, mas não resolveram o problema da recuperação e integração destes dados. Isso acontece porque estes serviços focam apenas as questões sintáticas referentes ao acesso, sem descrever detalhes sobre a semântica dos dados oferecidos pelo serviço. Recentemente, as infraestruturas de dados espaciais (IDE) (WILLIAMSON et al., 2003) têm conquistado uma grande popularidade como a solução para estas questões. Uma IDE pode ser definida como uma base relevante de tecnologias, políticas e acordos institucionais que facilitam a disponibilidade e acesso a dados espaciais, oferecendo uma base para descoberta, avaliação e aplicação para usuários e provedores de todos os níveis de governo, do setor comercial, do setor sem fins lucrativos, da academia e por cidadãos em geral (GSDI, 2004).

Desde a sua proposição, várias iniciativas para a criação de IDEs estão sendo desenvolvidas em todo o mundo. Alguns exemplos importantes destas iniciativas são a NSDI (National Spatial Data Infrastructure) (FGDC, 2005), nos Estados Unidos, o

Inspire (BERNARD et al., 2005), que é uma iniciativa européia, a ASDI (Australian Spatial Data Infrastructure) (ANZLIC Spatial Data Infrastructure Committee, 2004), de grupos australianos e neozelandeses, e a GSDI (HOLLAND et al., 1999), uma iniciativa que propõe a criação de uma IDE de nível global. No Brasil, também está sendo desenvolvida uma IDE de nível nacional, chamada de INDE (Infraestrutura Nacional de Dados Espaciais) (CONCAR, 2010).

Para facilitar a localização de informações, as IDEs atuais normalmente oferecem um serviço de catálogo, que é usado tanto por seus provedores de dados geográficos quanto pelos seus clientes. Os provedores usam este catálogo para anunciar o seu conjunto de dados. Para isto, eles realizam o cadastro do seu serviço, fornecendo uma série de informações que descrevem o seu conjunto de dados. Exemplos de informações que podem ser passadas durante este processo incluem o nome do provedor, uma descrição textual, a extensão geográfica coberta pelos dados, além de informações sobre o processo de produção dos dados. Todas estas informações são armazenadas em um registro de metadados, que é usado pelo catálogo durante o processamento de consultas. Por sua vez, os clientes da IDE usam o catálogo para localizar os dados geográficos do seu interesse. Sempre que uma consulta é realizada, o usuário pode avaliar cada registro recuperado para avaliar, dentre o conjunto de serviços selecionados, aqueles que melhor satisfazem as suas necessidades.

Apesar do desenvolvimento dos serviços de catálogo, a recuperação de dados geográficos nas IDEs atuais ainda é uma tarefa difícil de ser realizada. As próximas subseções descrevem alguns dos problemas das infraestruturas atuais, destacando as principais limitações que ocorrem para a recuperação de dados com base nas três dimensões que caracterizam dados geográficos: espaço, tema e tempo.

1.1.1 Consultas espaciais

Geralmente, consultas realizadas em portais geográficos requisitam a localização de mapas referentes a uma região geográfica específica. Exemplos de consultas espaciais incluem "*Encontre mapas sobre o estado da Paraíba*" ou "*Encontre mapas sobre o Nordeste do Brasil*". Nos serviços de catálogos atuais, a extensão geográfica coberta por um determinado serviço é representada através de um *bounding-box – menor retângulo envolvente*, que representa o menor retângulo possível, com lados paralelos aos eixos, que cobre toda a sua extensão geográfica. Além destas informações, alguns provedores incluem, entre as palavras-chave que descrevem o

serviço, os nomes das regiões cobertas pelo seu conjunto de dados. Além disto, durante a resolução de consultas espaciais, o usuário pode especificar o menor retângulo ou a geometria da região geográfica de seu interesse e um relacionamento topológico, tal como interseção, interno ou cobre, entre outros. Com base nestas informações, o catálogo recupera todos os serviços cuja extensão geográfica satisfaz os critérios definidos na requisição.

A primeira limitação na resolução de consultas espaciais nas IDEs atuais está relacionada à quantidade de informação que é repassada pelo provedor no momento em que o serviço é registrado no catálogo da infraestrutura. Normalmente, a maior parte dos provedores de dados geográficos usa um único registro para descrever todo o conjunto de dados oferecido pelo serviço. Nestas situações, uma única extensão geográfica é usada para representar todos os seus *feature types*, mesmo que frequentemente estas camadas se refiram a regiões geográficas diferentes. Quando este tipo de situação acontece, alguns problemas dificultam a resolução de buscas espaciais.

Para compreender melhor alguns dos problemas ocorridos, vamos considerar a Figura 1.1. Esta figura mostra dois registros de um serviço de catálogo, chamados M_1 e M_2, que descrevem diferentes serviços de dados geográficos. No serviço descrito por M_1, existem camadas que cobrem as regiões geográficas B_1, B_2 e B_3. Neste registro, foi definido que a extensão geográfica do serviço é a região geográfica B_5, que corresponde ao menor retângulo que cobre as extensões geográficas de todos os seus *feature types*. Os relacionamentos topológicos entre as regiões geográficas referentes a estes *bounding-boxes* são mostrados na Figura 1.2.

Para compreender o primeiro tipo de problema gerado por estas descrições, vamos considerar uma consulta na qual o usuário procura por mapas sobre a região geográfica B_4. Neste caso, como a extensão do registro M_1 intersecta a região definida na consulta, o registro acaba sendo recuperado, mesmo sem oferecer nenhum *feature type* que intersecte esta região. Outro tipo de problema ocorre se o usuário enviar uma consulta por mapas que intersectam a região geográfica B_3. Neste tipo de situação, o registro M_2 deixa de ser recuperado, embora o seu serviço tenha uma camada que intersecta a região solicitada. Isto acontece porque o seu provedor usou como a extensão geográfica do serviço apenas a região que é associada à maior parte dos seus *feature types*.

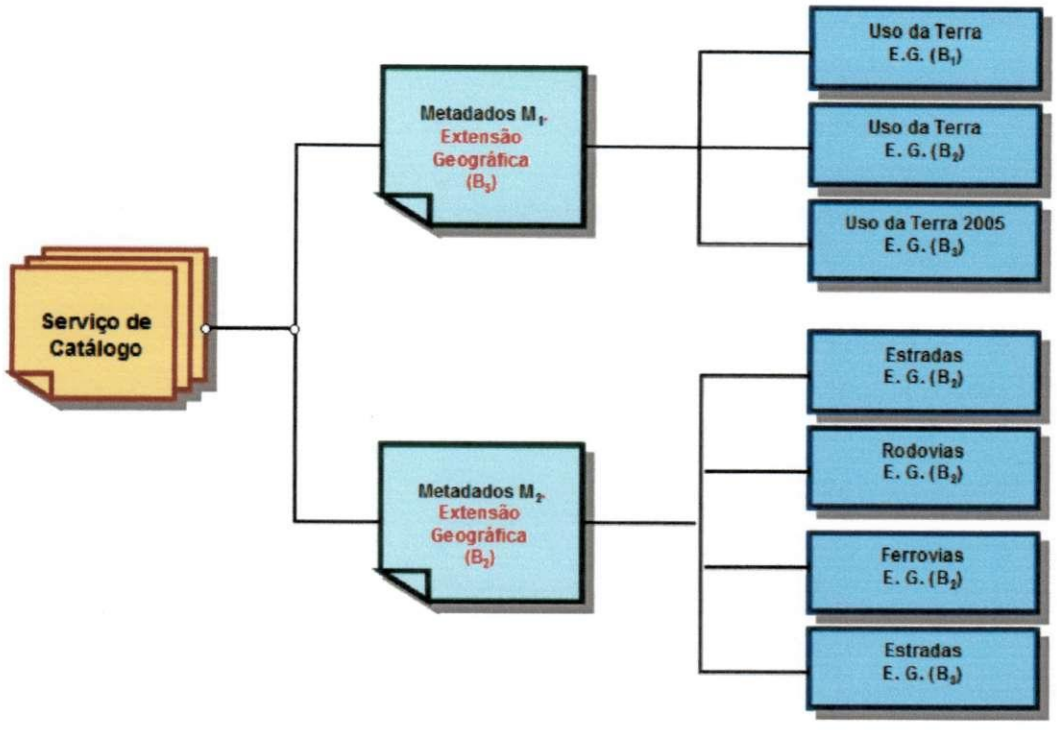

Figura 1.1: Exemplo de descrição geográfica de serviços

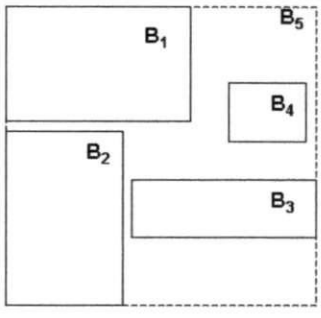

Figura 1.2: Relacionamentos topológicos entre bounding-boxes

Outra importante característica dos serviços de catálogos atuais é que os mesmos recuperam registros de metadados, que, na maior parte dos casos, descrevem o serviço como um todo. Desta forma, cabe ao usuário a tarefa de acessar o serviço selecionado e identificar, dentre todos os seus *feature types*, aqueles que satisfazem os critérios definidos na consulta. Tal tarefa, por muitas vezes, pode ser tediosa e consumir muito tempo, uma vez que muitos serviços oferecem uma grande quantidade de camadas. Ademais, a extensão geográfica de cada *feature type* é descrita através das coordenadas do seu *bounding-box*. Tal informação é difícil de ser avaliada visualmente pelo usuário.

Outra limitação importante dos serviços de catálogo atuais é que os mesmos consideram que todos os registros que satisfazem os critérios definidos na requisição

possuem a mesma relevância para o usuário. Neste caso, um registro que cobre apenas uma parte da região requisitada na consulta é considerado tão relevante quanto um registro que cobre totalmente esta região, e, por consequência, resolve completamente a consulta. O problema ocasionado por esta característica é que registros provavelmente mais relevantes podem ser mostrados tardiamente para o usuário durante a exibição do resultado de uma consulta, principalmente em requisições que retornam uma grande quantidade de registros. Além disto, em casos piores, estes serviços podem acabar nem sendo avaliados pelo usuário. Ou seja, um mecanismo de *ranking* encontra-se ausente.

1.1.2 Consultas temáticas

A dimensão tema é usada para identificar o tipo de camada que o usuário quer recuperar. Alguns exemplos possíveis de consultas com restrição temática são "*Encontre mapas sobre hidrografia*" ou "*Encontre mapas sobre reservas ambientais*". Nos serviços de catálogo atuais, a descrição do tema das camadas oferecidas por um serviço é realizada através de um conjunto de palavras-chave, que descrevem os principais temas relacionados ao mesmo. Ademais, informações sobre esta descrição podem ser encontradas em alguns atributos, como, por exemplo, em sua descrição textual.

Uma grande deficiência dos catálogos atuais para a resolução de consultas temáticas é que eles realizam as suas buscas apenas com base em palavras-chave, sem levar em consideração o significado da informação que está sendo requisitada e das informações usadas para descrever o serviço. Esta característica leva à realização de consultas com baixa cobertura, uma vez que serviços descritos com sinônimos ou termos relacionados às palavras-chave usadas na requisição deixam de ser recuperados.

Para entender este tipo de problema, vamos considerar os registros mostrados na Figura 1.3. Nesta figura, o serviço referente ao registro M_1 é descrito pelas palavras-chave *Rios*, *Açudes* e *Lagos*, enquanto o tema do serviço referente ao registro M_2 é descrito pela palavra-chave *Desastres Ambientais*. Neste exemplo, caso o usuário realize uma busca pelo tema *Corpos Hídricos*, o registro M_1 acaba não sendo recuperado, mesmo que todas as suas camadas sejam relacionadas ao tema definido na requisição. Já o registro M_2 é recuperado apenas se o usuário fizer uma busca pelo tema *Desastres Ambientais*. Assim, se o usuário fizer buscas por camadas mais específicas, como *Enchentes* ou *Queimadas*, o registro deixa de ser recuperado, mesmo oferecendo camadas referentes a estes temas.

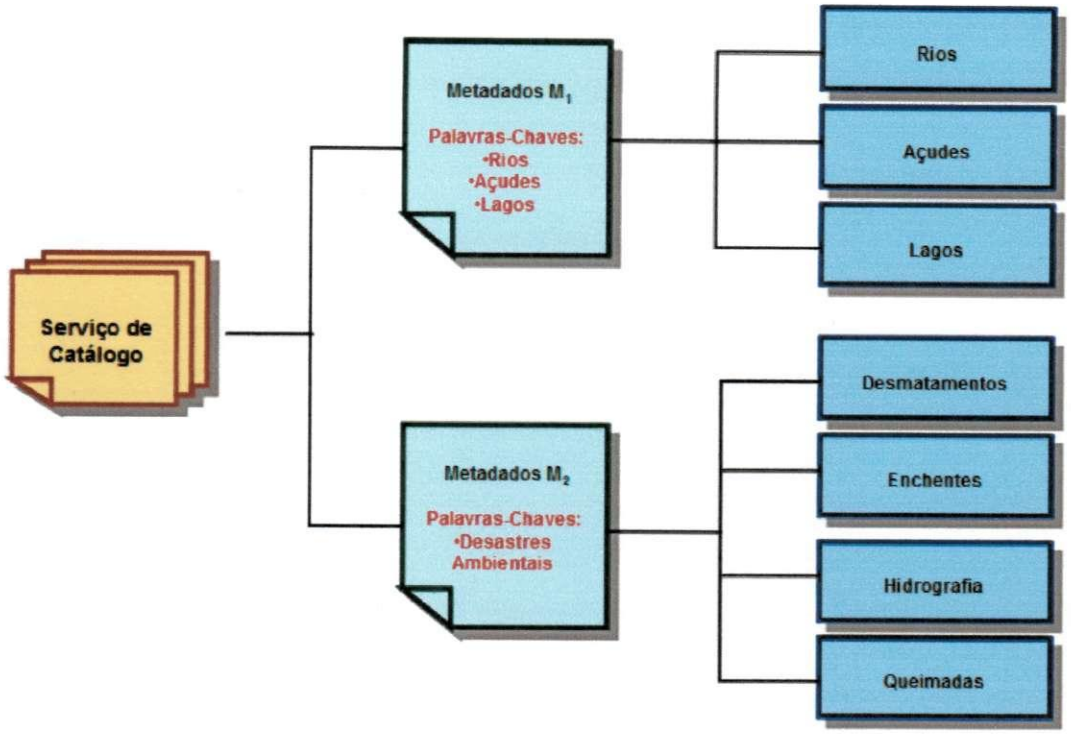

Figura 1.3: Exemplo de descrição temática de serviços

Uma forma de superar as limitações discutidas acima surgiu com o advento da *web* semântica (BERNERS-LEE et al., 2001). A *web* semântica corresponde a uma extensão da *web* atual, e o seu principal objetivo consiste na utilização de mecanismos que permitam descrever formalmente a semântica dos recursos publicados na rede. A implementação das ideias propostas pela *web* semântica requer a utilização de instrumentos que permitam descrever a semântica de um domínio de aplicação de uma forma que a mesma possa ser entendida tanto por humanos quanto por aplicações de *software*.

A implementação das ideias propostas pela web semântica é geralmente feita através de ontologias (GUARINO, 1995). Uma ontologia pode ser definida como a conceituação de um domínio de aplicação, enquanto que uma conceituação pode ser entendida como a forma de se pensar sobre um determinado domínio (USCHOLD, 1998). A vantagem do uso de ontologias é que elas permitem o desenvolvimento de agentes de *software* capazes de compreender o significado dos recursos publicados na *web*, o que melhora a qualidade das consultas e o compartilhamento de informações entre aplicações. Desde o seu surgimento, a *web* semântica tem conquistado uma grande popularidade, e é cada vez maior o número de aplicações, de diferentes domínios, que

usam suas ideias para melhorar o processo de recuperação de informação. Um domínio no qual estas ideias estão sendo comumente aplicadas é o geoespacial. Tal aplicação é chamada de *web* semântica geoespacial (EGENHOFER, 2002) (KUHN, 2005).

Após o advento da *web* semântica, vários trabalhos foram propostos abordando o uso de ontologias para melhorar a recuperação de dados geográficos. Nestes trabalhos, geralmente, os termos usados para a descrição dos recursos, assim como as palavras-chave utilizadas para a realização de consultas, consistem de conceitos pertencentes a uma ontologia. Nestes trabalhos, o processo de recuperação de dados normalmente acontece através de um relacionamento semântico chamado de subsunção (do inglês, *subsumption*). Neste tipo de consulta, são recuperados todos os recursos que são associados a conceitos que são subsumidos pelo conceito definido na requisição do usuário.

Assim como nas consultas realizadas pelos catálogos atuais, soluções baseadas em ontologias consideram que todos os recursos recuperados possuem a mesma relevância para o usuário. Desta forma, recursos que são associados exatamente ao conceito definido na consulta são considerados tão relevantes quanto aqueles que são associados, por exemplo, a conceitos que representam subclasses do mesmo, e que oferecem apenas uma parte das informações solicitadas. Isto também faz surgir a necessidade de uma medida de *ranking* que avalie a relevância de cada recurso recuperado para a consulta do usuário, considerando apenas a dimensão temática de ambos.

1.1.3 Consultas temporais

O tempo representa uma importante dimensão para muitas consultas que envolvem dados geográficos, como a recuperação de dados históricos, a geração de séries temporais de diversos temas, que permitem avaliar a evolução de um determinado fenômeno durante um período de tempo, e a recuperação de dados para a análise e o gerenciamento de desastres. Alguns exemplos possíveis de consultas com restrição temporal são "*Encontre mapas referentes ao ano de 2004*" ou "*Encontre mapas referentes ao período a partir de 2010*". Apesar da importância desta dimensão, a localização de camadas que satisfazem uma determinada restrição temporal ainda é uma tarefa difícil de ser realizada nas IDEs atuais, devido a algumas limitações dos seus serviços de catálogo.

Nos catálogos atuais, uma das informações requisitadas durante o registro de um novo serviço é a sua extensão temporal, que descreve o período de tempo referente aos dados que o mesmo oferece. A forma como esta informação é descrita pode variar de acordo com o padrão de metadados adotado pela IDE. No padrão ISO 19115 (ISO, 2003), que é o padrão especificado pela ISO para a documentação de dados geográficos e que serve como base para o desenvolvimento de vários outros padrões, a referência temporal de um serviço é geralmente descrita através de um intervalo. Tal intervalo é definido através de dois atributos que definem, respectivamente, os seus limites inicial e final.

Os catálogos atuais permitem a realização de consultas temporais com base em restrições definidas para o intervalo temporal associado ao serviço, ou através de restrições impostas sobre as datas de criação e modificação do registro que descreve o serviço. Entretanto, assim como nas dimensões espacial e temática, a resolução de consultas temporais nos catálogos atuais também apresenta uma série de limitações.

Alguns dos problemas que limitam a resolução de consultas temporais nos catálogos atuais são ilustrados na Figura 1.4, que mostra três registros de metadados, chamados M_1, M_2 e M_3. Cada registro descreve um serviço de dados geográficos, que oferece camadas referentes a diferentes intervalos temporais.

Uma grande limitação dos catálogos atuais ocorre porque muitas vezes, os valores dos atributos que descrevem a extensão temporal do serviço são omitidos pelo provedor no momento do registro. Este tipo de problema é mostrado no registro M_1. O serviço descrito por este registro oferece duas camadas referentes ao ano de 2005. Entretanto, como a extensão temporal do serviço foi omitida durante o registro do serviço e apresenta um valor nulo, estas camadas não podem ser recuperadas por qualquer consulta temporal. Um exemplo importante que ilustra bem este tipo de situação pode ser encontrado na IDE norte-americana [FGDC, 2005]. Uma análise de um conjunto de registros de metadados referentes a serviços de dados geográficos mostrou que quase 36% dos registros analisados não ofereciam esta informação. Em muitos destes casos, as únicas informações temporais oferecidas para estes serviços eram as datas em que o registro foi criado e/ou atualizado, que dificilmente descrevem a sua referência temporal com precisão.

A segunda limitação ocorre devido a problemas de consistência entre a extensão temporal definida no registro de metadados e as extensões temporais referentes às camadas oferecidas pelo serviço. Observando-se novamente a Figura 1.4, é possível

notar que o registro M_3 define como sua extensão temporal o intervalo 2005, embora também possua um *feature type* referente ao período de 2003. Este tipo de situação ocorre porque muitos provedores, no momento do cadastro do serviço, usam como valor da extensão temporal apenas o período que é associado à maior parte das suas camadas. Neste caso, se o usuário realizar uma consulta por mapas referentes ao período de 2003, o serviço de catálogo não vai recuperar o serviço descrito por M_3, apesar do mesmo possuir um *feature type* que satisfaz os critérios definidos na requisição.

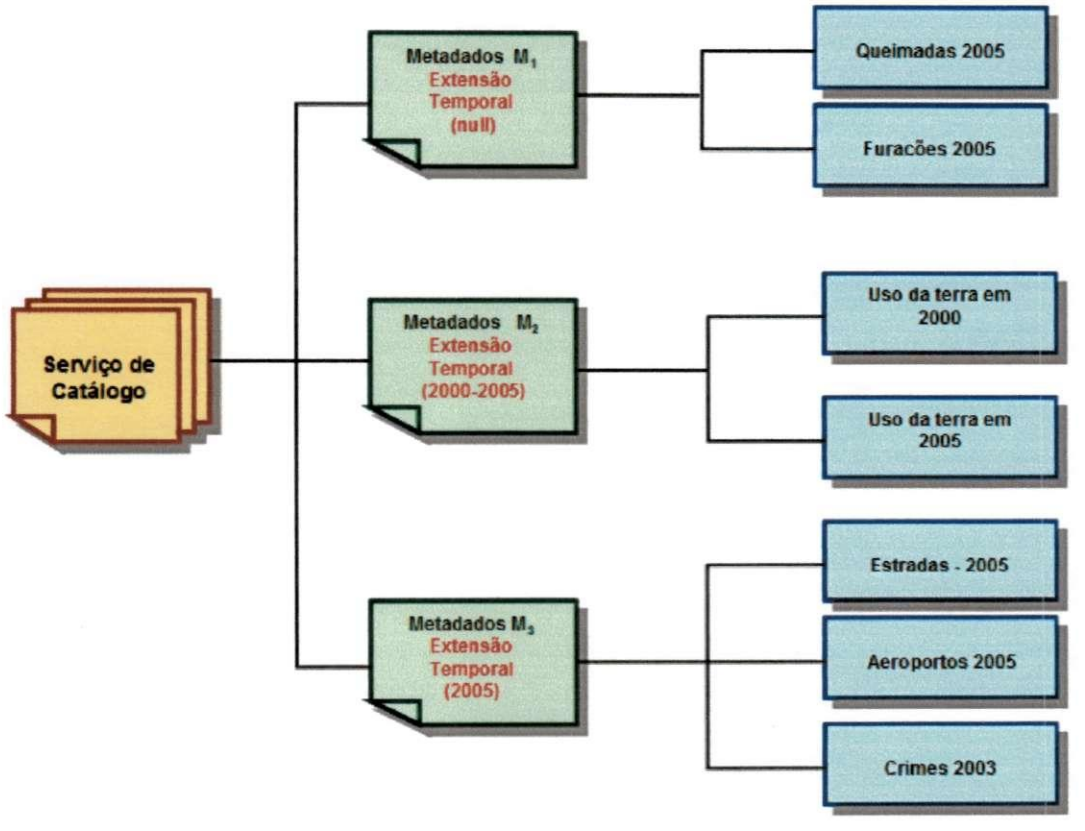

Figura 1.4: Exemplos de descrição temporal de serviços

Além das limitações descritas acima, os catálogos atuais consideram que todos os recursos que satisfazem o critério de seleção definido na consulta possuem a mesma relevância para o usuário. Desta forma, um serviço que cobre todo o intervalo temporal solicitado pelo usuário é considerado tão relevante quanto aquele que cobre apenas uma parte do intervalo solicitado. Esta limitação faz surgir a necessidade de uma métrica que permita avaliar o quanto um serviço oferecido pela infraestrutura é relevante para a consulta do usuário, considerando apenas a dimensão temporal de ambos.

1.1.4 Consultas globais

Em muitas situações, o cliente da infraestrutura pode estar interessado em dados geográficos que satisfazem restrições referentes a mais de uma dimensão. Nesta tese, este tipo de consulta é chamado de consulta global. Exemplos de consultas globais incluem "*Encontre mapas que mostrem os rios da Paraíba*, "*Encontre mapas sobre crimes ocorridos no ano de 2005*", "*Encontre mapas do Nordeste do Brasil entre os anos de 2000 e 2010*" e "*Encontre mapas sobre enchentes ocorridas no Rio de Janeiro no ano de 2011*". Este tipo de consulta, que é bastante frequente durante a recuperação de dados geográficos, é o mais difícil de ser resolvido pelos serviços de catálogos atuais. Esta dificuldade acontece porque este tipo de consulta combina as limitações inerentes a todas as dimensões envolvidas na consulta. Tal característica faz surgir a necessidade de se desenvolver mecanismos para facilitar a resolução deste tipo de consulta.

1.2 Objetivos e contribuições

O objetivo desta tese é melhorar o processo de recuperação de dados geográficos oferecidos por infraestruturas de dados espaciais. A resolução das limitações discutidas na seção anterior permite facilitar a localização dos dados geográficos disponíveis, aumentando a possibilidade de reutilização destas informações por parte dos clientes da infraestrutura. A principal contribuição desta pesquisa é o desenvolvimento de um arcabouço para facilitar a recuperação de dados geográficos oferecidos por IDEs, que oferece contribuições para as comunidades de bancos de dados, sistemas de informação geográfica e recuperação da informação. Outras contribuições relevantes propostas por este trabalho são:

- o desenvolvimento de um modelo de recuperação da informação geográfica para melhorar a descoberta de dados espaciais, que considera informações em nível de serviço e de *feature types*;
- uma nova abordagem para a anotação temática de *feature types* geográficos;
- o desenvolvimento de uma medida de *ranking* que permite avaliar a relevância de cada *feature type* oferecido pela infraestrutura para uma consulta do usuário, considerando apenas a dimensão espacial de ambos;
- o desenvolvimento de uma medida de *ranking*, baseada em ontologias, que permite avaliar a relevância de cada *feature type* oferecido pela

infraestrutura para uma consulta do usuário, considerando apenas a dimensão temática de ambos;

- o desenvolvimento de uma medida de *ranking* que permite avaliar a relevância de cada *feature type* oferecido pela infraestrutura para uma consulta do usuário, considerando apenas a dimensão temporal de ambos;
- o desenvolvimento de uma medida de *ranking* que permite avaliar a relevância de cada *feature type* oferecido pela infraestrutura para uma consulta do usuário, considerando duas ou mais dimensões.

É importante ressaltar que a solução proposta por esta tese não substitui o serviço de catálogo, que continua sendo responsável por armazenar a maior parte das informações sobre os serviços oferecidos pela IDE. Entretanto, a sua função é facilitar a localização dos dados geográficos oferecidos por este catálogo. Por este motivo, sempre que o arcabouço proposto realiza uma consulta, oferece referências que ligam cada camada recuperada ao seu respectivo registro no serviço de catálogo, no qual o cliente pode encontrar informações mais detalhadas sobre como o dado foi produzido, informações de qualidade, restrições de uso, entre outras.

1.3 Hipóteses do trabalho

Para o desenvolvimento desta tese, foram levantadas as seguintes hipóteses:

- o desenvolvimento de um modelo de recuperação da informação geográfica, que descreva as características espaciais, temáticas e temporais de cada *feature type* pode facilitar a localização dos dados oferecidos pelo catálogo da infraestrutura;
- o uso de ontologias pode melhorar consideravelmente a cobertura e a precisão de consultas temáticas;
- é possível identificar as características espaciais, temáticas e temporais de forma automática a partir das informações já disponíveis, sem gerar novos esforços para o provedor durante o processo de cadastro de serviços;

- as ideias aplicadas à modelagem de documentos na recuperação da informação clássica podem ser reaproveitadas para melhorar o *ranking* dos recursos recuperados.

1.4 Visão geral

A Figura 1.5 mostra uma visão geral da solução proposta da tese, destacando o processo de coleta de informações. A primeira etapa deste processo consiste em obter informações junto ao serviço de catálogo da infraestrutura. O objetivo desta etapa é recuperar os registros que descrevem novos serviços de dados geográficos oferecidos pelo catálogo. Ao fim desta etapa, para cada registro recuperado, é obtido um documento XML contendo uma série de informações acerca do conjunto de dados oferecido pelo serviço, incluindo a URL na qual o mesmo pode ser acessado.

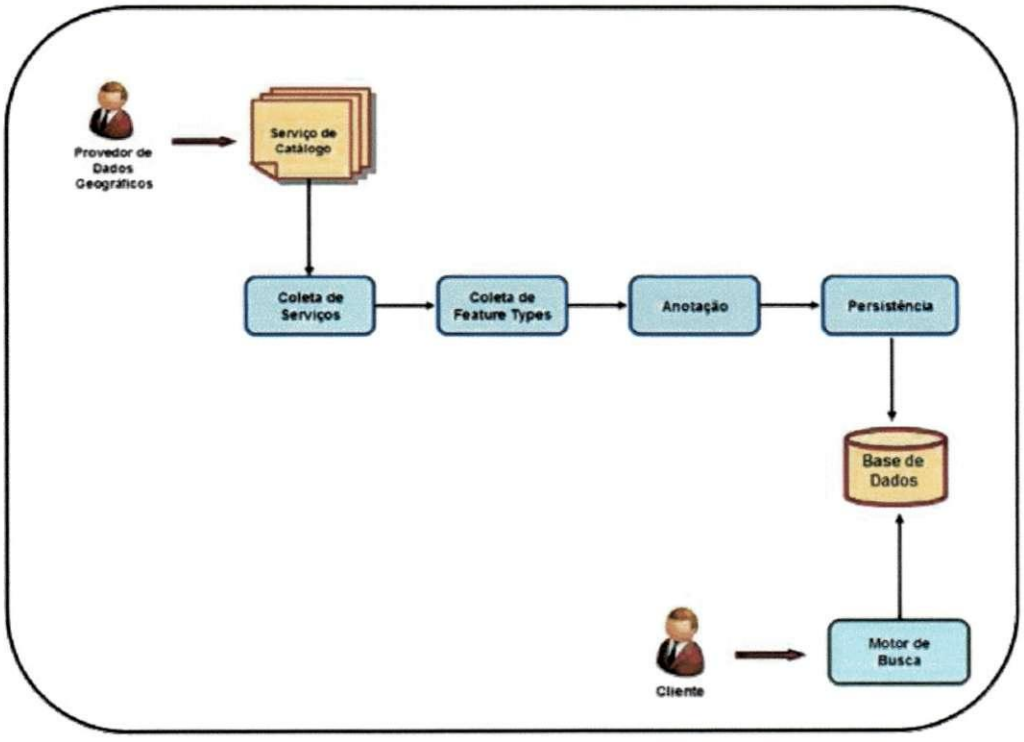

Figura 1.5: Visão geral da solução proposta

Conforme discutido anteriormente, o registro de metadados recuperado junto ao catálogo contém várias informações acerca do serviço como um todo, mas não possui informações detalhadas a respeito dos *feature types* que o mesmo oferece. Por esta

razão, a segunda etapa consiste em acessar o serviço junto ao seu respectivo provedor. O objetivo desta etapa é obter o documento de funcionalidades (do inglês *capabilities*) do serviço, que contém informações referentes a cada uma de suas camadas.

A terceira etapa consiste na anotação de cada *feature type* oferecido pelo serviço. Nesta etapa, as informações obtidas a partir do registro de metadados e do documento de funcionalidades são usadas para identificar as informações referentes a cada camada. Para cada camada, são extraídos: a sua extensão geográfica, as suas palavras-chave e a sua extensão temporal. Visando melhorar a qualidade das consultas temáticas, as palavras-chave identificadas para descrever cada *feature type* se referem a conceitos definidos em ontologias. É importante ressaltar que todas estas informações são identificadas de forma automática, e não requerem qualquer esforço adicional do provedor no momento em que o serviço é incluído no catálogo da infraestrutura.
Finalmente, a última etapa do processo de coleta de informações consiste em armazenar e indexar as informações obtidas e produzidas ao longo das etapas anteriores em um banco de dados. Uma vez persistidas, estas informações tornam-se disponíveis para a ferramenta de busca do arcabouço, que possui uma série de *matchmakers* que permitem a resolução de consultas com restrições espaciais, temáticas, temporais e globais, que possuem dois ou três tipos de restrição.

1.5 Organização do documento

O restante deste documento está organizado como segue. No capítulo 2, é apresentada uma visão geral sobre os conceitos e tecnologias que foram usados para o desenvolvimento da tese. No capítulo 3, é contemplado o levantamento bibliográfico, revisando os principais trabalhos propostos na literatura referente à área de concentração da pesquisa. O capítulo 4 apresenta a solução proposta pela tese. O capítulo 5 descreve os processos de coleta de informações. O capítulo 6 descreve o processo de recuperação da informação. O capítulo 7 descreve a avaliação experimental. Por fim, o capítulo 8 conclui a tese, mostrando as conclusões obtidas e os próximos trabalhos referentes à continuação da pesquisa.

Capítulo 2 – Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica da tese, oferecendo uma visão geral sobre os principais conceitos e as tecnologias utilizados para a sua elaboração. É importante salientar que não é objetivo do capítulo oferecer uma discussão aprofundada sobre cada assunto. Entretanto, para cada tema abordado, são dadas referências para locais nos quais uma discussão mais detalhada sobre o mesmo pode ser encontrada. Primeiramente, o capítulo aborda a recuperação da informação clássica, voltada para descoberta de documentos. Depois, são discutidos a *web* semântica, ontologias, padrões de metadados espaciais e serviços OGC. Por fim, é oferecida uma visão geral sobre infraestruturas de dados espaciais e arquiteturas orientadas a serviços.

2.1 Recuperação da informação clássica

A recuperação da informação é uma área de pesquisa antiga na ciência da computação. A recuperação da informação clássica (BAEZA-YATES; RIBEIRO-NETO, 1999) é caracterizada pelo desenvolvimento de soluções voltadas para a localização de documentos. Em seus modelos, tanto os documentos disponíveis quanto as consultas de usuários são representados como uma coleção de termos ou palavras-chave. O processo de localização de informação, por sua vez, consiste em recuperar, a partir de uma ou mais palavras-chave, os documentos que podem ser relevantes para o usuário. Com base nestas ideias, vários modelos foram desenvolvidos. Contudo, três deles são considerados clássicos e servem como base para o desenvolvimento de outros modelos. Estes modelos são: o modelo booleano, o modelo vetorial e o modelo probabilístico.

No modelo booleano (SALTON, 1989), os documentos publicados e as consultas do usuário são representadas como vetores n-dimensionais, sendo n o número de termos que podem ser usados para a indexação de documentos. A principal característica deste modelo é que cada dimensão destes vetores só pode receber valores binários. No vetor que representa a consulta do usuário, para cada dimensão que representa um termo presente na consulta do usuário é atribuído o valor 1, enquanto o valor 0 é atribuído para as demais dimensões. Já no vetor que representa um documento, para cada dimensão que representa um termo presente no documento é atribuído o valor 1, enquanto o valor 0 é atribuído para as demais dimensões. O processamento de uma

consulta consiste em recuperar, dentre o conjunto de documentos existentes, apenas aqueles cujo vetor é idêntico ao vetor que representa a consulta do usuário. A principal vantagem deste modelo é a facilidade de implementação. Contudo, sua principal desvantagem é que não há como estabelecer um *ranking* entre os documentos recuperados. Outra desvantagem importante é que o mesmo não permite o casamento parcial de documentos. Desta forma, documentos que casam parcialmente com a consulta são julgados irrelevantes e não são recuperados para o usuário.

No modelo vetorial (SALTON; LESK, 1968) (SALTON, 1971), os documentos e as consultas também são representados como vetores n-dimensionais. Assim como no modelo booleano, cada dimensão desses vetores representa um termo usado para a indexação dos documentos. Contudo, no modelo vetorial as dimensões destes vetores podem receber valores não binários. Normalmente, quando este modelo é aplicado, o peso de cada dimensão é determinado através do produto de duas métricas. A primeira delas é chamada de *tf* (do inglês, *term frequency*) e representa a frequência com que o termo representado pela dimensão ocorre dentro do documento. A segunda é chamada de *idf* (do inglês, *inverse document frequency*) e representa a frequência com que este mesmo termo aparece na coleção de documentos do sistema.

Depois que os valores das frequências *tf* e *idf* são computados, o grau de relevância de um documento para a consulta do usuário pode ser calculado através de medidas como o cosseno do ângulo formado pelo vetor da consulta e o vetor do documento, ou a distância euclidiana entre os pontos que representam estes vetores. Durante a realização de uma consulta, todos os documentos que possuem um grau de relevância superior a um *threshold* mínimo são recuperados. A grande vantagem deste tipo de modelo é que o cálculo de uma medida de similaridade permite que os documentos sejam classificados e ordenados de acordo com sua relevância, permitindo que documentos provavelmente mais relevantes sejam apresentados primeiro para o usuário. Ademais, o modelo também permite a recuperação de documentos que satisfazem parcialmente a consulta do usuário. A desvantagem do modelo vetorial é que o mesmo não considera a correlação existente entre os termos.

Por fim, o modelo probabilístico (ROBERTSON; SPARCK JONES, 1976) é outro modelo bastante utilizado para a recuperação de documentos. Como o próprio nome sugere, este modelo usa a teoria das probabilidades para mensurar o quanto um documento é relevante para uma consulta formulada por um usuário. Neste modelo, dois conjuntos são formados, sendo que um contém os documentos que provavelmente são

relevantes para o usuário, enquanto o outro armazena os que provavelmente não são relevantes. Durante uma consulta, o sistema de busca calcula a probabilidade de cada documento pertencer a cada um destes conjuntos, e a proporção entre estas probabilidades é usada para estabelecer o grau de relevância de cada documento. Futuras interações do usuário com os resultados apresentados podem ser usadas para redefinir estas probabilidades e realizar sucessivos refinamentos. Enquanto este modelo permite o *ranking* de documentos com base no valor de suas probabilidades, a sua principal desvantagem consiste na dificuldade em definir os critérios que devem ser adotados para a realização do primeiro refinamento.

A principal limitação dos sistemas clássicos de recuperação da informação é a utilização de palavras-chave para a indexação e recuperação de documentos. Essa característica compromete tanto a cobertura quanto a precisão das suas consultas. A cobertura diminui porque documentos relevantes que possuem em seu conteúdo termos que representam sinônimos ou outros termos que são relacionados aos termos usados na consulta acabam não sendo recuperados. A precisão diminui porque documentos não relevantes que possuem o termo definido na consulta em sua composição acabam sendo recuperados para o usuário.

Outra característica importante da *web* atual é que, embora a maior parte do seu conteúdo esteja na forma de documentos, outros tipos de conteúdo coexistem. Exemplos dessas novas formas de conteúdo incluem dados multimídia, como arquivos de imagem, áudio e vídeo e dados geográficos. Estes tipos de dados são ditos semi ou não estruturados, o que faz surgir a necessidade de uma nova forma de encontrar e manter estes dados com facilidade. Para complementar, o advento de tecnologias como arquitetura orientada a serviços e *web services* fez com que a Internet também hospede uma grande quantidade de serviços, que podem ser descobertos e invocados através da rede. Uma vez que é difícil descrever serviços com palavras-chave, outros meios são necessários para a descoberta deste tipo de recurso.

2.2 A *Web* semântica

A *web* semântica foi proposta para resolver as limitações dos sistemas clássicos de recuperação da informação. Esta nova *web* foi proposta como uma extensão da Internet atual, na qual os recursos oferecidos são anotados com metadados que descrevem o seu significado, tornando-os compreensíveis tanto para humanos quanto para máquinas. Por meio desta extensão, é possível descrever formalmente os

relacionamentos semânticos existentes entre os recursos disponibilizados na rede, possibilitando uma melhor localização de recursos, e permitindo a automatização de uma grande quantidade de tarefas. Fensel et al. (2003) propõem uma arquitetura para a implementação da *web* semântica, a qual pode ser vista como uma pirâmide composta por vários elementos. Esta arquitetura é mostrada na Figura 2.1, que foi extraída de (FENSEL et al., 2003):

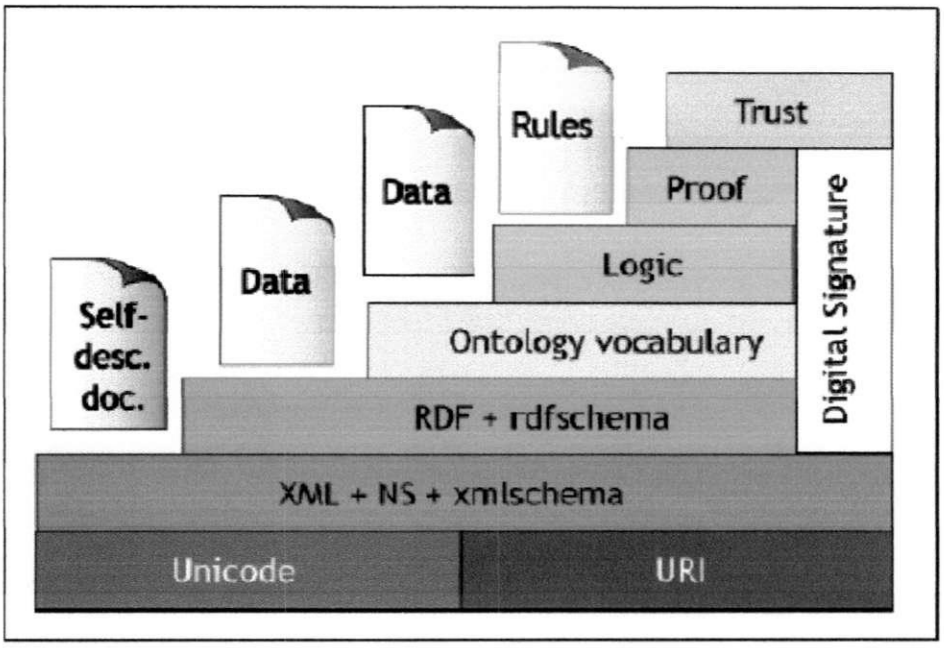

Figura 2.1: Pilha de elementos da web semântica, extraída de (FENSEL et al., 2003)

A primeira camada corresponde aos recursos publicados na rede, através da URI usada para a sua identificação e dos caracteres usados para descrever o seu conteúdo. Logo acima desta camada, está a linguagem XML, que representa um padrão aberto para troca de informações na *web*. Esta linguagem permite uma maior interoperabilidade entre as aplicações, uma vez que os seus documentos têm uma forma livre, ou seja, não precisam obedecer a uma estrutura pré-definida. No entanto, quando necessário, pode-se aplicar alguma estrutura a estes documentos, o que pode ser feito de duas formas. A primeira forma é através de um DTD, que usa uma gramática para definir a estrutura dos elementos e a ordem em que os mesmos devem ser aninhados dentro do documento XML. A segunda forma é através de um esquema definido na linguagem *XML Schema* (W3C, 2001), na qual a estrutura é definida por meio de um documento XML.

A terceira camada é representada pela linguagem RDF (*Resource Description Framework*) (W3C, 1999), utilizada para fazer assertivas em relação aos recursos publicados. As construções RDF são baseadas em triplas do tipo objeto-atributo-valor. Uma tripla deste tipo é representada da forma A (O, V), na qual A representa o atributo, O representa o objeto, e V representa o valor deste atributo para o objeto. Uma sentença RDF pode ser lida como *"O objeto O tem um atributo A com o valor V"*. Também nesta camada, está a linguagem RDF-Schema (W3C, 2004a), que pode ser utilizada para a construção de ontologias simples. Esta linguagem permite a definição de classes, atributos e relacionamentos entre as classes. Além disto, ela também oferece construtores para a definição de subclasses e subpropriedades.

A quarta camada contém as ontologias usadas durante os processos de descrição e descoberta de informações. Para a definição de ontologias, foram desenvolvidas linguagens que possuem um poder de expressividade maior do que RDFS. Exemplos de linguagens para a definição de ontologias são DAML (DAML, 1999), OIL (FENSEL et al., 2001), *Description Logic* (BAADER et al., 2003) e OWL (W3C, 2004b).

A próxima camada corresponde à lógica, que permite a descrição de conhecimento a partir de conceitos e regras de restrição que podem derivar taxonomias de classificação automaticamente. Esta lógica pode ser usada também para a inferência de novos conhecimentos durante o processo de descoberta de informações.

As camadas de prova e confiança são usadas para a validação do conhecimento obtido a partir da camada de lógica, avaliando a qualidade e a confiabilidade da informação obtida.

2.3 Ontologias

Para implementar as ideias propostas para a *web* semântica, é necessário criar instrumentos que permitam descrever formalmente a semântica dos domínios de aplicação referentes aos recursos que são publicados na rede. Isto vem sendo resolvido com o uso de ontologias.

A palavra ontologia vem da filosofia e representa a ciência de descrever as entidades do mundo e seus relacionamentos. Considerando-se a aplicação deste termo na ciência da computação, pode-se dizer que uma ontologia define os termos utilizados para descrever e representar uma determinada área de conhecimento. Assim, é possível criar ontologias para diferentes domínios de aplicação, como relevo, hidrografia, vegetação, sistemas de transportes, atividades econômicas, entre outros. As informações

descritas em ontologias podem ser utilizadas como uma referência por pessoas e aplicações que precisam localizar e compartilhar informações na rede.

Em uma ontologia, a semântica de um domínio de conhecimento é descrita a partir dos seguintes componentes:

- **classes:** representam as entidades do domínio que está sendo descrito, ou seja, tudo aquilo sobre o que se quer descrever informações;
- **propriedades:** definem o conjunto de atributos que caracterizam cada entidade;
- **relacionamentos:** representam os relacionamentos semânticos existentes entre as entidades do domínio, como, por exemplo, generalização, agregação, composição, disjunção, entre outros;
- **instâncias:** representam o conjunto de objetos de cada classe definida na ontologia. Cada instância corresponde a uma concretização de uma classe definida na ontologia;
- **axiomas:** uma das principais características de ontologias é que elas permitem a definição de axiomas. Estes axiomas são definidos por meio de regras, e são usados para a inferência de novos conhecimentos durante o processamento de uma consulta.

Uma vez criadas, ontologias podem ser usadas por ferramentas de busca para melhorar a qualidade de suas consultas. Quando um processo de descoberta de informação é baseado em ontologias, dois tipos de conhecimento podem ser identificados. O primeiro deles é chamado de conhecimento *intensional*, e representa o conhecimento que é obtido através da análise das classes que compõem a ontologia. O segundo tipo de conhecimento é chamado de *extensional* e é obtido através da análise das instâncias associadas à ontologia.

Ao longo dos anos, várias linguagens foram propostas para o desenvolvimento de ontologias, como DAML e OIL, enquanto outras linguagens também foram usadas para este fim, como RDF-S. Atualmente, duas linguagens ocupam um destaque maior, sendo as mais utilizadas para a realização deste tipo de tarefa: *Description Logic* e OWL.

Description Logic é uma linguagem baseada em lógica, na qual os componentes da ontologia são definidos através de um conjunto de regras. A linguagem é bastante

expressiva, e oferece construtores para definir relacionamentos semânticos e restrições de integridade. Exemplos de construções oferecidas por esta linguagem incluem composição, disjunção, negação e quantificadores existenciais. Em *Description Logic*, uma ontologia contém dois tipos de *framework*. O primeiro é chamado de *T-Box*, e contém a parte da definição dos conceitos que compõem a ontologia, enquanto o segundo é chamado de *A-Box*, e contém o conjunto de instâncias pertencentes à ontologia.

Por sua vez, OWL representa a linguagem recomendada pelo W3C para a criação de ontologias. Uma ontologia OWL pode descrever classes, propriedades e instâncias de uma forma bastante simples. É baseada nas linguagens XML e RDF, e usa os esquemas destas duas linguagens na definição de suas estruturas. OWL também provê expressividade para a definição de ontologias, mas não oferece construções para a especificação de axiomas. Contudo, outras linguagens para a definição de regras, como Rule ML e SWRL, podem ser usadas para a realização desta tarefa. A linguagem OWL subdivide-se em três tipos, de acordo com o nível de expressividade oferecido:

- ***OWL Lite:*** é o tipo mais simples de OWL, no qual todos os relacionamentos entre classes têm cardinalidade 0 ou 1;
- ***OWL DL:*** representa um tipo mais elaborado, que permite um poder de representação preservando a decidibilidade computacional. Desta forma, tudo que é representado através desta linguagem pode ser entendido e decidido por uma máquina. É o tipo mais utilizado.
- ***OWL Full:*** representa o tipo de OWL que tem o maior poder de representação, embora não se tenha garantia de que o que ela representa seja computacionalmente decidível. Um exemplo do poder de representação de *OWL Full* é que nela, uma classe pode assumir o papel de instância em alguns momentos, e vice-versa.

2.4 Arquitetura orientada a serviços

Recentemente, arquiteturas orientadas a serviços (do inglês - *Service Oriented Architecture* - SOA) (ERL, 2005) têm se tornando uma importante solução para permitir a interoperabilidade entre aplicações na Internet. Neste tipo de arquitetura, as funcionalidades oferecidas por um sistema são encapsuladas e oferecidas como um conjunto de serviços que podem ser invocados remotamente através da rede. Ademais, esses serviços possuem uma interface aberta e a troca de mensagens entre aplicações

normalmente acontece através de protocolos baseados em XML, o que permite que um serviço seja invocado sem a necessidade de se conhecer detalhes internos sobre a sua implementação e a plataforma pelo qual o mesmo é oferecido.

Em SOA, três tipos de componentes podem ser identificados: provedores de serviço, clientes e serviços de diretório. Os provedores de serviço usam a rede para oferecer um ou mais serviços. Os clientes, por sua vez, utilizam os serviços oferecidos por estes provedores. Para que os clientes possam localizar serviços, existe a necessidade de uma estrutura que forneça informações sobre os serviços que estão disponíveis. Essa tarefa é realizada pelo serviço de diretório, que é usado tanto por provedores quanto por clientes. Provedores usam o diretório para anunciar os seus serviços, enquanto que os clientes o utilizam para descobrir os serviços nos quais eles estão interessados.

Uma característica importante de SOA é que os serviços de diretório não são responsáveis pelo armazenamento dos serviços oferecidos. Entretanto, sempre que uma consulta é realizada, o diretório retorna para os clientes um conjunto de informações que descrevem cada serviço recuperado. Com base nessas informações, o cliente pode avaliar esses serviços, identificar aquele que melhor atende as suas necessidades e acessá-lo junto ao seu respectivo provedor. Toda a interação entre os componentes de SOA é mostrada na Figura 2.2.

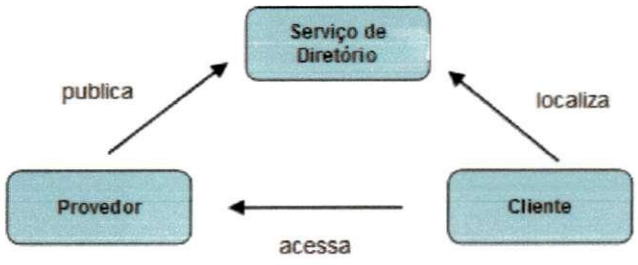

Figura 2.2: Interação entre os componentes em uma arquitetura SOA

Um dos fatores que contribuíram para a popularização de SOA foi o desenvolvimento da tecnologia de *web services* (ALONSO, G., et al., 2004). Essa tecnologia define um conjunto de padrões para a implementação destas arquiteturas. Entre estes padrões, estão o protocolo SOAP, que é um protocolo baseado em XML para a troca de mensagens entre aplicações, a linguagem WSDL para a descrição de serviços e o padrão UDDI para a implementação de serviços de diretórios. Outro estilo

para a implementação de SOA, chamado REST, foi proposto por Fielding (2000). Neste estilo, serviços podem ser invocados diretamente através do protocolo HTTP, sem a necessidade de utilização de SOAP. Serviços implementados usando esse estilo de intereação são chamados de *RESTful web services*.

Desde a sua proposição, arquiteturas orientadas a serviços têm sido bastante utilizadas em diversos domínios de aplicação, incluindo o domínio geoespacial. A popularidade de SOA, aliada à grande aceitação e utilização dos serviços propostos pelo OGC, tem feito com que este tipo de arquitetura tenha sido cada vez mais utilizada no processo de implementação das infraestruturas de dados espaciais mais recentes (BERNARD; CRAGLIA, 2005).

2.5 Padrões de metadados espaciais

Para facilitar a sua descoberta e reutilização, os dados geográficos disponíveis precisam ser documentados. Durante esse processo de anotação, são disponibilizadas informações acerca dos dados que estão sendo oferecidos. Através destas informações, usuários podem avaliar se os dados satisfazem ou não as suas necessidades. Um dos grandes problemas enfrentados com relação à documentação de dados geográficos é que esse processo era feito de forma heterogênea, uma vez que as documentações existentes podiam conter diferentes tipos de informação e eram normalmente podiam ser oferecidas em diferentes formatos. Visando superar estas limitações, vários padrões para a descrição de metadados referentes a dados geográficos começaram a ser desenvolvidos.

Um padrão bastante conhecido para a geração de metadados é o *Dublin Core* (DC, 2005), que é composto por um conjunto de quinze elementos usados para a documentação de recursos oferecidos na *web*. Este padrão, que não é voltado para qualquer domínio de aplicação em particular, tornou-se bastante popular e foi usado por muitos provedores de informação. Contudo, o suporte limitado para a descrição das características geoespaciais dos recursos disponibilizados e a simplicidade de seus elementos, que são representados apenas por uma descrição textual, são algumas das características que limitam a sua aplicação no domínio de dados geográficos.

Outro padrão bastante popular é o CSDGM (*Content Standard for Digital Geospatial Metadata*) (FGDC, 1998), que é um padrão norte-americano proposto em 1994 e revisado em 1998 pelo FGDC. Diferentemente do padrão *Dublin Core*, esta norma foi desenvolvida especificamente para a descrição de informações geográficas.

Este padrão conquistou uma grande aceitação e, além dos Estados Unidos, passou a ser utilizado em diversos locais, como Canadá, Reino Unido e África do Sul. Outros padrões de metadados espaciais foram propostos por organizações européias, como os projetos *LaCLEF* (EUROGI, 1999) e ESMI (SCHOLTEN, 1998).

Visando unificar a especificação de metadados no domínio de dados geográficos, a ISO, através do seu comitê técnico ISO TC 211, desenvolveu o padrão ISO 19115 (ISO, 2003). Este padrão define uma série de elementos que devem ser utilizados para a descrição de informações geográficas, sendo complementado por outros padrões especificados pelo mesmo comitê, como o ISO 19139 (ISO, 2007), que define como os elementos do padrão ISO 19115 devem ser codificados em um arquivo XML, e o ISO 19108 (ISO, 2002), que define um conjunto de elementos para a descrição das características temporais.

A norma ISO 19115 é bastante ampla, sendo composta por cerca de quatrocentos elementos. Os seus elementos permitem detalhar várias características dos dados que estão sendo descritos, como a sua identificação, as suas propriedades espaciais e temporais, proveniência, informações de qualidade, entre outras. Este padrão também oferece flexibilidade, uma vez que nem todos os elementos precisam ser adotados.

Outra característica importante desta norma é que a mesma permite que os seus elementos sejam estendidos. Desta forma, cada usuário deste padrão tem liberdade para usá-lo da forma mais conveniente para si, definindo os elementos que serão adotados e até mesmo criando seus próprios elementos. Cada "adaptação" do padrão é chamada de perfil. O padrão define ainda uma série de elementos que são considerados como pertencentes a um núcleo. Estes elementos devem estar presentes em cada perfil definido para a norma.

Desde a sua proposição, esta norma tem conquistado uma popularidade muito grande, e está cada vez mais se tornando um padrão de fato entre os produtores de dados geográficos. Uma prova disto é que a mesma está sendo adotada como o padrão de metadados geográficos em grande parte das IDEs que estão sendo desenvolvidas no mundo todo. Uma IDE que está adotando o padrão ISO 19115 é a INDE, cujo perfil para o padrão é chamado de MGB (Metadados Geoespaciais do Brasil) (CEMG, 2009). O perfil foi submetido a uma consulta pública e será adotado por todos os provedores que pretendam tomar parte da IDE.

2.6 Serviços OGC

Para resolver os problemas de heterogeneidade referentes ao domínio de dados geográficos, foi criado, em 1994, o OGC (*Open Geospatial Consortium*), que é um consórcio formado por diversos tipos de organização, como agências de governos, instituições acadêmicas e empresas privadas. O principal objetivo deste consórcio é o desenvolvimento de padrões públicos e abertos para o domínio geoespacial, visando aumentar a interoperabilidade e o reuso de dados geográficos.

Desde a sua criação, o OGC já desenvolveu uma série de padrões. Dentre estes padrões, estão uma série de serviços que podem ser usados para descobrir e acessar dados geográficos oferecidos em diversos tipos de formato. Estes serviços são especificados na forma de *web services*, ou seja, para cada serviço proposto, é especificada apenas a sua interface, contendo as operações que devem ser implementadas e seus respectivos parâmetros de entrada e saída, sem descrever os detalhes de sua implementação. Isto diminui o acoplamento entre os serviços e facilita a interoperabilidade entre aplicações. As próximas seções discutem alguns dos principais serviços propostos pelo OGC: o *Web Map Service* (WMS), o *Web Feature Service* (WFS) e o *Catalog Service for the Web* (CSW).

2.6.1 Web Map Service (WMS)

O WMS é o serviço proposto pelo OGC para a descoberta e recuperação de dados vetoriais para apresentação. O mapa gerado pelo serviço é transmitido para o usuário geralmente na forma de um arquivo de imagem, embora o formato SVG (W3C, 2002) também seja permitido. Este serviço oferece três tipos de operações: *GetCapabilities*, *GetMap* e *GetFeatureInfo*. As duas primeiras são obrigatórias, enquanto a última é opcional.

A operação *GetCapabilities* pode ser usada pelo cliente para descobrir, além de informações gerais a respeito do serviço, quais são as camadas oferecidas pelo mesmo. O seu resultado é um arquivo XML contendo informações gerais a respeito da configuração e das funcionalidades do serviço, além da descrição de todos os seus *feature types*. Para cada camada oferecida pelo serviço, são descritas informações como o nome que deve ser usado para a sua recuperação, o seu título, a sua descrição textual, as palavras-chave que a descrevem, e a região geográfica coberta pela mesma. Após executar esta operação, o cliente pode analisar estas informações para avaliar as camadas nas quais ele tem interesse.

A Figura 2.3 mostra a invocação da operação *GetCapabilities* de um serviço WMS oferecido pela Agência Nacional de Águas, que é uma agência federal brasileira. Os parâmetros de entrada utilizados representam, respectivamente, o nome do serviço, a operação que deve ser executada e a versão do serviço que deve ser utilizada.

```
http://200.140.135.184/cgi-bin/mapserv?map=/usr/local/www/apache22/data/hidro
/wshidro.map
&SERVICE=WMS
&REQUEST=GetCapabilities
&VERSION=1.0.0
```

Figura 2.3: Exemplo de requisição GetCapabilities de um WMS

A Figura 2.4 mostra uma das camadas descritas pelo documento de funcionalidades obtido a partir da execução da operação *GetCapabilities* descrita pela URL da Figura 2.3. O nome da camada descrita, que deve ser usada para a sua recuperação, é *rios_federais*. Outras informações, como o título, descrição textual, palavras-chave e extensão geográfica também podem ser encontradas.

```
- <Layer queryable="1">
    <Name>rios_federais</Name>
    <Title>Base hidrorreferenciada - Rios Federais</Title>
    <Abstract>Rios federais</Abstract>
  - <Keywords>
      Hidrografia Rios Federais OTTO PFAFSTETTER
    </Keywords>
    <SRS>epsg:4291 epsg:4326</SRS>
    <LatLonBoundingBox minx="-79.5229" miny="-34.9285" maxx="-34.7896" maxy="5.7266"/>
    <BoundingBox SRS="EPSG:4291" minx="-79.5229" miny="-34.9285" maxx="-34.7896" maxy="5.7266"/>
  </Layer>
```

Figura 2.4: Descrição de uma camada no serviço WMS

Depois de recuperar e avaliar as camadas oferecidas pelo serviço, o cliente pode usar a operação *GetMap* para recuperar as camadas de seu interesse. A Figura 2.5 mostra um exemplo de requisição de mapa usando a operação *GetMap*, enviada ao mesmo serviço usado para a operação *GetCapabilities*. Neste exemplo, após a URL do serviço, são informados vários parâmetros de entrada, que definem respectivamente: o tipo de serviço que está sendo acessado, a versão que deve ser utilizada, a operação que deve ser executada, uma lista contendo os nomes das camadas que devem ser recuperadas, o sistema de coordenadas que deve ser adotado, as coordenadas da

extensão geográfica de interesse, a altura da imagem que deve ser gerada, a largura da imagem e o formato de saída.

```
http://200.140.135.184/cgi-bin/mapserv?map=/usr/local/www/apache22/da
ta/hidro/wshidro.map
      &SERVICE=WMS
      &VERSION=1.1.1
      &REQUEST=GetMap
      &LAYERS=rios_federais
      &CRS=EPSG:4326
      &BBOX=-79.5229,-34.9285,-34.7896,5.7266
      &HEIGHT=500&
      &WIDTH=800&
      &FORMAT=image/jpeg
```

Figura 2.5: Exemplo de uma requisição GetMap

Como saída, a operação *GetMap* retorna um arquivo de imagem referente ao mapa gerado a partir da sobreposição de todas as camadas solicitadas na requisição do usuário. A Figura 2.6 mostra o mapa obtido através da requisição da Figura 2.5.

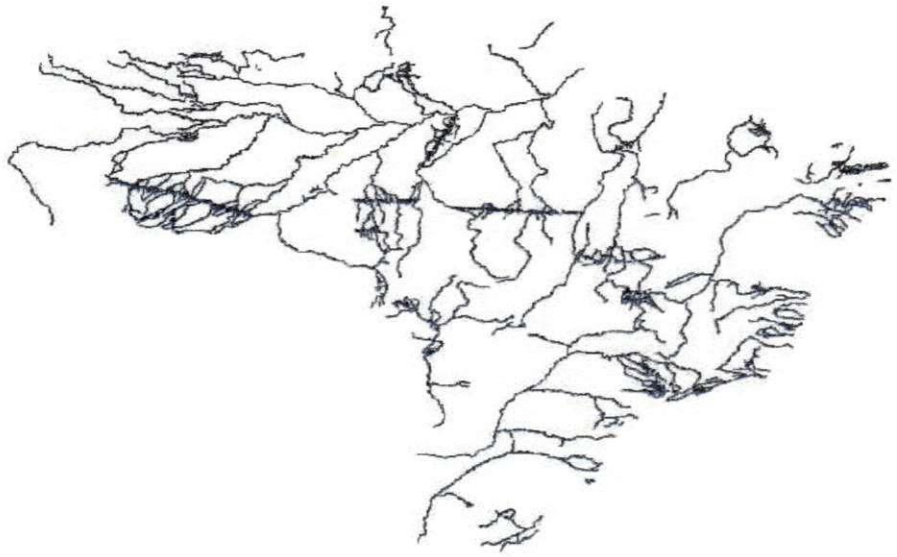

Figura 2.6: Mapa gerado pela operação GetMap

A operação *GetFeatureInfo* é usada para a recuperação de informações adicionais sobre feições mostradas no mapa gerado após a invocação da operação *GetMap*. A implementação desta operação é opcional para o provedor que está

oferecendo o serviço. Detalhes sobre a sua interface podem ser encontrados em (OGC, 2004).

2.6.2 Web Feature Service (WFS)

Assim como o serviço WMS, o serviço WFS também permite a recuperação de dados vetoriais. Contudo, neste tipo de serviço, estas informações são recuperadas em outros formatos, tais como GML, que é uma linguagem baseada em XML para a especificação de feições espaciais, e *shapefile*. Este serviço oferece operações que permitem tanto a recuperação quanto a atualização de informações junto ao servidor. As operações de atualização, no entanto, são disponíveis apenas para usuários autorizados. As principais operações deste serviço, que podem ser usadas para a descoberta e recuperação de informações, são: *GetCapabilities*, *DescribeFeatureType* e *GetFeature*.

A operação *GetCapabilities* é usada para descobrir as camadas oferecidas pelo serviço. O seu resultado é um documento XML contendo informações gerais sobre o serviço, além, de uma descrição de seus *feature types*. Para cada camada, são mostradas informações como o nome usado para a sua identificação e recuperação, o seu título, sua descrição textual, as palavras-chave que a descrevem e a região geográfica coberta pela mesma.

A Figura 2.7 mostra uma requisição para a operação *GetCapabilities* de um serviço WFS oferecido pelo Ministério do Meio Ambiente brasileiro. Os demais exemplos desta subseção também são baseados nesse serviço. Os parâmetros de entrada desta requisição representam, respectivamente, o nome do serviço, a operação que deve ser invocada e a versão do serviço que deve ser utilizada.

```
http://mapas.mma.gov.br/cgi-bin/mapserv?map=/opt/www/html/webservices
/biorregioes.map
      &service=WFS
      &request=GetCapabilities
      &version=1.0.0
```

Figura 2.7: Exemplo de requisição GetCapabilities do WFS

A Figura 2.8 mostra a descrição de um dos *feature types* do documento XML retornado pela execução do serviço a partir da URL da Figura 2.7. Por meio desta

figura, é possível perceber que as informações descritas para cada camada são parecidas com aquelas usadas em um documento de funcionalidades do serviço WMS.

```
<FeatureType>
  <Name>areas_priori_import</Name>
 - <Title>
    Revisão áreas prioritárias para conservação da biodiversidade (importância biológica)-2007
  </Title>
 - <Abstract>
    Revisão áreas prioritárias para conservação da biodiversidade (2007)-classificadas de acordo com a importância biológica
  </Abstract>
 - <Keywords>
    áreas prioritárias conservação biodiversidade revisão importância biológica
  </Keywords>
  <SRS>EPSG:4291</SRS>
  <LatLongBoundingBox minx="-73.9899" miny="-35.9627" maxx="-25.9996" maxy="7.0335"/>
</FeatureType>
```

Figura 2.8: Descrição de um tipo de feição em um serviço WFS

Depois de analisar as camadas oferecidas pelo serviço, o cliente pode então selecionar mais detalhes sobre as informações oferecidas por um *feature type*. Isto pode ser feito através da operação *DescribeFeatureType*. Esta operação é usada para recuperar o esquema usado para a descrição de uma ou mais camadas oferecidas pelo serviço. Para invocá-la, o usuário informa uma lista contendo o nome de todos os *feature types* que devem ser descritos. Caso o nome destas camadas sejam omitidos, o serviço retorna a descrição de todos os seus *feature types*.

A Figura 2.9 representa uma requisição usada para invocar a operação *DescribeFeatureType,* com o objetivo de recuperar a descrição do esquema da camada mostrada na Figura 2.8. Os parâmetros de entrada presentes na requisição representam o tipo de serviço utilizado, a operação que deve ser invocada, a versão do serviço e a lista de *feature types* que devem ser descritos.

```
http://mapas.mma.gov.br/cgi-bin/mapserv?map=/opt/www/html/webservices
/biorregioes.map
      &SERVICE=WFS
      &REQUEST=DescribeFeatureType
      &VERSION=1.0.0
      &TYPENAME=areas_priori_import
```

Figura 2.9: Exemplo de requisição DescribeFeatureType

O resultado da operação *DescribeFeatureType* é um documento XML contendo as descrições dos esquemas dos *feature types* solicitados. Cada descrição recuperada é apresentada através de um *XML Schema,* e contém informações como o nome, o tipo e a cardinalidade de cada atributo que a caracteriza. A Figura 2.10 mostra o esquema obtido a partir da invocação solicitada pela URL da Figura 2.9.

```
- <complexType name="areas_priori_importType">
  - <complexContent>
    - <extension base="gml:AbstractFeatureType">
      - <sequence>
          <element name="msGeometry" type="gml:GeometryPropertyType" minOccurs="0" maxOccurs="1"/>
          <element name="gid" type="string"/>
          <element name="acao_prior" type="string"/>
          <element name="bioma" type="string"/>
          <element name="sub_bioma" type="string"/>
          <element name="nome" type="string"/>
          <element name="cod_id" type="string"/>
          <element name="tipo" type="string"/>
          <element name="importancia" type="string"/>
          <element name="prioridade" type="string"/>
          <element name="acao1" type="string"/>
          <element name="acao2" type="string"/>
          <element name="acao3" type="string"/>
          <element name="acao4" type="string"/>
          <element name="acao5" type="string"/>
          <element name="acao6" type="string"/>
          <element name="cria_uc" type="string"/>
          <element name="grupo_uc" type="string"/>
          <element name="caracter" type="string"/>
          <element name="oportun" type="string"/>
          <element name="ameacas" type="string"/>
        </sequence>
      </extension>
    </complexContent>
```

Figura 2.10: Descrição do esquema de um tipo de feição no serviço WFS

Outra operação importante do serviço WFS é a *GetFeature*, que pode ser usada para recuperar os objetos associados a um ou mais *feature types*. Para invocar esta operação, o usuário informa o nome de todos os tipos de feições de interesse, e o serviço recupera todas as feições pertencentes a cada tipo solicitado. O resultado é normalmente um arquivo GML contendo as feições de cada camada requisitada. Entretanto, outros formatos, como *shapefile*, também podem ser utilizados durante esta operação.

Por definição, a operação *GetFeature* recupera todas as feições associadas a cada *feature type* solicitado. Entretanto, existem situações nas quais o usuário está interessado em apenas uma parte destas feições, de acordo com um determinado critério

de seleção. Nestes casos, o usuário pode definir uma ou mais restrições de filtros, que são aplicadas ao conjunto de objetos que estão sendo recuperadas. As restrições de filtro podem ser especificadas através de um filtro OGC (OGC, 2005b). Estes filtros permitem a utilização de operadores espaciais (e.g., *Equals, Disjoint*, *Intersect*, *Touches*) para selecionar objetos de acordo com relacionamentos topológicos inerentes a sua geometria, além de operadores escalares para a seleção de objetos de acordo com os valores de seus atributos descritivos. Quando restrições de filtros são especificadas durante esta operação, o serviço retorna para o usuário, apenas as feições que satisfazem as restrições definidas por ele.

A Figura 2.11 mostra como as feições do *feature type* mostrado no exemplo anterior podem ser recuperadas. Os parâmetros de entrada presentes na requisição representam, respectivamente, o tipo de serviço utilizado, a operação que deve ser invocada, a versão do serviço e a lista de feições que devem ser recuperadas. Neste caso, como nenhuma restrição de filtro foi especificada e o número de resultados não foi limitado, todas as feições dessa camada são recuperadas.

```
http://mapas.mma.gov.br/cgi-bin/mapserv?map=/opt/www/html/webservices
/biorregioes.map
      &SERVICE=WFS
      &REQUEST=GetFeature
      &VERSION=1.0.0
      &TYPENAME=areas_priori_import
```

Figura 2.11: Exemplo de requisição GetFeature

A Figura 2.12 mostra a descrição de uma feição recuperada a partir da invocação da Figura 2.11. Através da mesma, pode-se perceber que as informações são descritas de acordo com o *XML Schema* obtido a partir da operação *DescribeFeatureType*.

O serviço WFS oferece também outras operações para os seus usuários. A operação *GetGMLObject* permite que feições GML sejam recuperadas diretamente através de uma identificação. A operação *Transaction* pode ser utilizada pelo provedor para gerenciar os dados oferecidos pelo serviço. Para isto, a operação oferece opções que permitem a inclusão de novos dados, bem como a atualização e a exclusão dos dados existentes. Para evitar problemas de consistência durante o processo de atualização, a operação *LockFeature* pode ser usada para bloquear os objetos que estão sendo

modificados. Detalhes sobre o funcionamento de todas estas operações podem ser encontrados em (OGC, 2005c).

```
-<gml:featureMember>
 -<ms:areas_priori_import>
  +<gml:boundedBy></gml:boundedBy>
  +<ms:msGeometry></ms:msGeometry>
   <ms:gid>2672</ms:gid>
   <ms:acao_prior>Orden Pesq e Área Excl Pesca</ms:acao_prior>
   <ms:bioma>Zona Marinha</ms:bioma>
   <ms:sub_bioma/>
   <ms:nome>Talude continental</ms:nome>
   <ms:cod_id>Zm012</ms:cod_id>
   <ms:tipo/>
   <ms:importancia>Muito Alta</ms:importancia>
   <ms:prioridade>Extremamente Alta</ms:prioridade>
  -<ms:acao1>
     Proposição de áreas de exclusão de pesca (arrasto principalmente) em áreas de ocorrência conhecida ou potencial de recifes profundos
   </ms:acao1>
   <ms:acao2>Ordenamento pesqueiro</ms:acao2>
  -<ms:acao3>
     Monitoramento em áreas de interface da ocorrência de recifes profundos com quaisquer atividades de exploração de recursos que seja potencialmente impactantes
   </ms:acao3>
   <ms:acao4/>
   <ms:acao5/>
   <ms:acao6/>
   <ms:cria_uc>Não</ms:cria_uc>
   <ms:grupo_uc/>
  +<ms:caracter></ms:caracter>
  -<ms:oportun>
     Existência de estudos e potencial de parcerias; estruturas de produção como atratores de grandes pelágicos; efeito restritivo a pesca nas áreas onde se localizam estruturas de produção
   </ms:oportun>
  +<ms:ameacas></ms:ameacas>
  </ms:areas_priori_import>
</gml:featureMember>
```

Figura 2.12: Feição recuperada por um serviço WFS

2.6.3 O serviço de catálogo (CSW)

Muitas vezes, é comum que uma determinada fonte de informação ofereça aos seus usuários, uma série de dados e serviços geográficos. Surge então a necessidade de um serviço que permita que clientes possam descobrir, de forma simples, quais são os serviços oferecidos por um provedor e como eles podem ser acessados. Para oferecer esta funcionalidade, o OGC desenvolveu um serviço de catálogo. Este serviço atua como um registro, que pode ser usado tanto por provedores de dados geoespaciais quanto por clientes. Provedores usam o catálogo para anunciar os seus dados e serviços, através da publicação dos metadados que descrevem estes recursos. Os clientes, por sua vez, podem usar o catálogo para recuperar e analisar as informações publicadas neste serviço. Após esta descoberta, os usuários podem então analisar os registros obtidos para verificar se os recursos que os mesmos descrevem satisfazem as suas necessidades. Caso o usuário encontre algum recurso de seu interesse, ele pode usar as informações

contidas no registro de metadados para acessá-lo junto ao seu provedor correspondente. O serviço de catálogo proposto pelo OGC oferece uma série de operações para os seus clientes. As operações mais importantes para a localização de informações junto ao serviço são *GetCapabilities*, *GetRecords* e *GetRecordById*.

A operação *GetCapabilities* é usada para descobrir as funcionalidades do serviço de catálogo. O seu resultado é um documento XML contendo uma série de informações acerca do serviço, como o seu título, sua descrição textual, as informações descritivas sobre o seu provedor, palavras-chave e as operações que são implementadas. A Figura 2.13 mostra como pode ser solicitada a recuperação do documento de funcionalidades oferecido pelo serviço de catálogo do INSPIRE, que é a IDE européia. Os parâmetros de entrada representam, respectivamente, o nome do serviço, a operação que deve ser executada e a versão que deve ser utilizada. A Figura 2.14 mostra um trecho do documento de funcionalidades recuperado através deste serviço, referente à descrição do provedor do serviço de catálogo.

```
http://www.inspire-geoportal.eu/discovery/csw?
      SERVICE=CSW
      &REQUEST=GetCapabilities
      &VERSION=2.0.2
```

Figura 2.13: Exemplo de requisição GetCapabilities do CSW

A localização de informações no serviço de catálogo é realizada através da operação *GetRecords,* que pode ser invocada para recuperar os registros de metadados que são oferecidos pelo serviço. Para executá-la, o cliente deve definir uma série de restrições, como o tipo de recurso no qual está interessado (conjunto de dados, serviços, aplicações ou coleção de conjuntos de dados), o número máximo de registros que devem ser recuperados durante a execução e o padrão de metadados que deve ser usado para a exibição dos registros. Contudo, todas estas restrições precisam estar de acordo com as opções oferecidas pelo serviço. As informações acerca destas opções podem ser facilmente encontradas no seu documento de funcionalidades.

A Figura 2.15 mostra o código XML usado para realizar uma requisição para a operação *GetRecords* de um serviço de catálogo. Para diminuir a quantidade de

informação que será recuperada, o elemento *ElementSetName* é usado para que cada registro recuperado seja descrito de forma simplificada.

```
- <ows:ServiceProvider>
    <ows:ProviderName>INSPIRE geoportal catalogue service</ows:ProviderName>
    <ows:ProviderSite xlink:href="http://www.inspire-geoportal.eu"/>
  - <ows:ServiceContact>
      <ows:IndividualName>Hildegard Gerlach</ows:IndividualName>
      <ows:PositionName>Reviewer</ows:PositionName>
    - <ows:ContactInfo>
      - <ows:Phone>
          <ows:Voice/>
          <ows:Facsimile/>
        </ows:Phone>
      - <ows:Address>
          <ows:DeliveryPoint>TP 262</ows:DeliveryPoint>
          <ows:City>Ispra</ows:City>
          <ows:AdministrativeArea>Italy</ows:AdministrativeArea>
          <ows:PostalCode>21027</ows:PostalCode>
          <ows:Country>it</ows:Country>
          <ows:ElectronicMailAddress>ioannis.kanellopoulos@jrc.ec.europa.eu</ows:ElectronicMailAddress>
        </ows:Address>
        <ows:HoursOfService/>
        <ows:ContactInstructions>EC Joint Research Centre</ows:ContactInstructions>
      </ows:ContactInfo>
      <ows:Role>uni</ows:Role>
    </ows:ServiceContact>
  </ows:ServiceProvider>
```

Figura 2.14: Identificação do provedor em um serviço de catálogo

```
- <csw:GetRecords service="CSW" version="2.0.2" resultType="results">
  - <csw:Query typeNames="csw:Record">
      <csw:ElementSetName>brief</csw:ElementSetName>
    </csw:Query>
  </csw:GetRecords>
```

Figura 2.15: Exemplo de requisição GetRecords

A Figura 2.16, por sua vez, mostra três dos registros obtidos quando esta solicitação é enviada para o serviço de catálogo oferecido pelo INSPIRE. Como foi solicitada apenas uma breve descrição dos registros, para cada resultado recuperado, são mostrados apenas o seu identificador, o título, o tipo de informação descrita pelo registro e a sua extensão geográfica.

A operação *GetRecordById* permite que o usuário do serviço recupere um ou mais registros de metadados de acordo com suas identificações. Tal tarefa pode ser usada quando o cliente já sabe antecipadamente quais registros ele quer acessar. Esta operação também pode ser usada como um complemento para a operação *GetRecords*.

Por exemplo, para não oferecer muita informação ao usuário, a operação *GetRecords* pode recuperar apenas uma breve descrição para cada registro, e depois a operação *GetRecordById* pode ser usada para recuperar mais informações sobre registros nos quais o usuário está de fato interessado.

```
- <csw:BriefRecord>
    <dc:identifier>jrc-img2k_pr1_uk12_pan</dc:identifier>
    <dc:title>Image2000 Product 1 (uk12) Panchromatic</dc:title>
    <dc:type>dataset</dc:type>
  - <ows:BoundingBox
    crs="::PROJCS["British_National_Grid",GEOGCS["GCS_OSGB_1936",DATUM["D_OSGB_1936",SPHEROID["A
      <ows:LowerCorner>-0.18 54.89</ows:LowerCorner>
      <ows:UpperCorner>-4.12 56.92</ows:UpperCorner>
    </ows:BoundingBox>
  </csw:BriefRecord>
- <csw:BriefRecord>
    <dc:identifier>jrc-img2k_pr1_fr16_pan</dc:identifier>
    <dc:title>Image2000 Product 1 (fr16) Panchromatic</dc:title>
    <dc:type>dataset</dc:type>
  - <ows:BoundingBox
    crs="::PROJCS["France_IMAGE2000",GEOGCS["GCS_NTF",DATUM["D_NTF",SPHEROID["Clarke_1880_IGN
      <ows:LowerCorner>4.59 49.29</ows:LowerCorner>
      <ows:UpperCorner>1.19 51.27</ows:UpperCorner>
    </ows:BoundingBox>
  </csw:BriefRecord>
- <csw:BriefRecord>
    <dc:identifier>jrc-img2k_pr1_se16_pan</dc:identifier>
    <dc:title>Image2000 Product 1 (se16) Panchromatic</dc:title>
    <dc:type>dataset</dc:type>
  - <ows:BoundingBox
    crs="::PROJCS["Swedish_National_Grid_RT_1990",GEOGCS["GCS_RT_1990",DATUM["D_RT_1990",SPHERO
      <ows:LowerCorner>17.27 61.80</ows:LowerCorner>
      <ows:UpperCorner>12.30 63.90</ows:UpperCorner>
    </ows:BoundingBox>
  </csw:BriefRecord>
```

Figura 2.16: Resultado da operação GetRecords

A Figura 2.17 mostra o código XML usado para realizar uma requisição para a operação *GetRecordById,* com o objetivo recuperar mais informações a respeito de um dos registrados mostrados na Figura 2.16. A requisição indica, através do parâmetro *outputSchema*, que o registro deve ser descrito de acordo com o padrão de metadados espaciais ISO 19115. O resultado desta operação é um documento XML contendo a descrição completa do registro solicitado, no formato de metadados especificado na requisição.

O serviço de catálogo CSW ainda oferece outras operações. A operação *DescribeRecord* pode ser usada para obter uma descrição do esquema que mostra os elementos usados para a descrição dos registros. A operação *GetDomain* permite que

um cliente recupere informações a respeito do domínio dos elementos que compõem o registro de metadados ou a respeito dos parâmetros usados nas requisições. A operação *Transaction* permite que novos metadados sejam cadastrados no serviço de catálogo, além da atualização e exclusão dos registros existentes. A operação *Harvest* pode ser usada para a recuperação de registros durante um processo de atualização. Detalhes sobre a definição destas operações podem ser encontrados em (OGC, 2007a).

```
- <csw:GetRecordById service="CSW" version="2.0.2" outputSchema="http://www.isotc211.org/2005/gmd">
    <csw:Id>jrc-img2k_pr1_uk12_pan</csw:Id>
  </csw:GetRecordById>
```

Figura 2.17: Requisição para a operação GetRecordById

2.7 Infraestruturas de dados espaciais

Uma infraestrutura de dados espaciais (IDE) pode ser definida como uma base relevante de tecnologias, políticas e acordos institucionais para facilitar a disponibilidade e o acesso a dados geoespaciais (GSDI, 2004). Desde a sua proposição, as IDEs têm conquistado uma grande popularidade para garantir a interoperabilidade entre dados geográficos produzidos por diferentes organizações, facilitando a sua disseminação, o seu acesso, e, consequentemente, a sua reutilização.

O desenvolvimento de uma IDE envolve vários tipos de participantes, e, por isto, depende da resolução de muitas questões de ordem técnica, política e institucional. Segundo Warnest (2005), este desenvolvimento está fundamentado sobre cinco pilares:

- **normas e padrões:** um dos principais objetivos das IDEs consiste em garantir a interoperabilidade de dados geográficos produzidos por diferentes provedores. Isto é alcançado através da definição de normas e padrões que devem ser seguidos por todas as fontes de dados que vão usar a infraestrutura para oferecer os seus dados e aplicações. Exemplos de normas e padrões incluem o sistema de coordenadas que deve ser usado para a representação dos dados; o padrão de metadados que deve ser usado para a documentação; os modelos de dados que devem ser usados para a disponibilização dos dados e os serviços que devem ser adotados para oferecer acesso aos dados;

- **instituições:** representam os acordos e as articulações institucionais necessários para a implementação de uma IDE, incluindo as políticas que devem ser adotadas por cada empresa, a legislação que vai regulamentar estes acordos, a coordenação entre os participantes, entre outros. Estes acordos também resolvem questões como a custódia e o licenciamento dos dados;
- **tecnologia:** representa toda a infraestrutura necessária para garantir o acesso, a distribuição e o armazenamento dos dados, tais como bases de dados, equipamentos de *hardware*, infraestrutura de rede, *softwares*, entre outros;
- **dados:** representam todos os conjuntos de dados geográficos oferecidos por fontes de dados que compõem a infraestrutura. Atualmente, por questões de interoperabilidade, o acesso a estes dados é normalmente oferecido através de serviços padronizados, como, por exemplo, o WMS e o WFS;
- **atores:** correspondem a todo o conjunto de clientes que interagem com a IDE. Exemplos de atores incluem os provedores de informação geográfica, que usam a infraestrutura para anunciar os seus dados e serviços, e clientes, que a utilizam para localizar e acessar as informações geográficas de seu interesse.

Como uma IDE precisa prover a localização de dados oferecidos por diversas agências, são necessários mecanismos para permitir que novos provedores de dados espaciais possam anunciar os dados e serviços que oferecem. Durante este processo, o usuário responsável pelo registro da fonte de dados precisa fornecer uma série de informações acerca dos seus recursos, de acordo com o padrão de metadados adotado pela IDE. Depois que o cadastro é finalizado, os metadados informados durante esse processo são armazenados e utilizados para a resolução de consultas.

Uma interface para implementação do serviço de registro é o serviço de catálogo especificado pelo OGC. A grande vantagem do uso deste serviço é que ele permite que o processo de publicação de metadados seja realizado programaticamente através de uma aplicação de *software*. No entanto, quando metadados são adicionados a uma infraestrutura, existe uma grande preocupação em assegurar que as suas informações estejam de acordo com os padrões adotados pela mesma. Tal preocupação visa garantir

a interoperabilidade dos recursos oferecidos pela IDE. Mohammadi et al. (2008), propõem a implementação de um serviço responsável por verificar se os recursos adicionados à infraestrutura são interoperáveis com os demais recursos oferecidos pela mesma.

Outra forma bastante comum de permitir o registro de metadados é através de uma interface gráfica, disponibilizada através de portais geográficos. Neste caso, o usuário responsável por realizar o registro preenche uma série de formulários, que repassam informações acerca dos seus recursos. Depois que o registro é finalizado, as informações recebidas são codificadas no formato do padrão de metadados da infraestrutura e armazenadas. A vantagem deste tipo de interface é que usuários não familiarizados com o serviço de catálogo podem realizar o cadastro de seus recursos.

Além de permitir a publicação de novas informações, uma infraestrutura também precisa oferecer mecanismos que permitam que os recursos publicados possam ser localizados pelos clientes. Da mesma forma que o serviço de registro, a interface de um serviço que permite a descoberta de informações espaciais em um servidor é oferecida pelo padrão CSW. Como esta interface é padronizada, clientes podem consultar várias infraestruturas de maneira uniforme. A grande vantagem da utilização deste serviço é que outras aplicações podem enviar consultas para a IDE. Contudo, para muitos usuários, não familiarizados com este tipo de serviço, as infraestruturas geralmente oferecem também portais geográficos (MAGUIRE; LONGLEY, 2005), que permitem que consultas sejam realizadas de forma mais simples através de interfaces gráficas.

2.8 Considerações finais

Este capítulo descreveu os principais conceitos e tecnologias utilizadas para a elaboração desta tese. O embasamento teórico oferecido pelo mesmo vai facilitar o entendimento dos próximos capítulos do documento. O próximo capítulo mostra como as tecnologias e conceitos apresentados aqui estão sendo usados para melhorar a recuperação de informações em infraestruturas de dados espaciais e portais geográficos.

Capítulo 3 – Revisão Bibliográfica

Nos últimos anos, vários trabalhos têm sido propostos para melhorar a localização de dados geográficos oferecidos por infraestruturas de dados espaciais e portais geográficos. Além de utilizarem diferentes abordagens para a resolução deste problema, os trabalhos propostos variam de acordo com o tipo de recurso que pode ser recuperado pelo cliente. Enquanto alguns trabalhos são mais genéricos e permitem a descoberta de qualquer tipo de recurso oferecido pela IDE, outros trabalhos focam na descoberta de recursos específicos, como os serviços de dados geográficos (WMS, WFS e WCS), serviços de processamento de dados geográficos, *feature types* e feições. As próximas seções discutem os trabalhos que mais se destacam nesta área de pesquisa. Para facilitar o entendimento do capítulo, os trabalhos foram agrupados de acordo com o tipo de recurso que é recuperado.

3.1 Recuperação de informação em nível de serviços

O problema da recuperação de serviços de processamento de dados espaciais consiste em localizar serviços que oferecem uma determinada funcionalidade desejada pelo usuário. A localização de serviços é um problema em aberto não só para a comunidade espacial, mas para toda a comunidade de recuperação da informação. A análise dos trabalhos propostos mostra que dois tipos de abordagens têm se destacado na resolução deste problema.

O primeiro tipo consiste em associar as características de cada serviço, como parâmetros de entrada e saída, precondições e efeitos, a conceitos definidos em ontologias. Este tipo de solução é geralmente implementado com a utilização de tecnologias desenvolvidas para a anotação semântica de serviços *web*, tais como WSMO (BRUIJN et al., 2005), DAML-S (ANKOLEKAR et al., 2001) e OWL-S (MARTIN et al., 2005). O segundo tipo de abordagem consiste na criação de uma ou mais ontologias para a classificação de serviços. Neste caso, a anotação semântica é realizada através da associação dos serviços disponíveis a classes definidas nestas ontologias.

3.1.1 O trabalho de Klien et al.

Klien et al. (2006) desenvolveram uma solução para a recuperação de serviços de processamento de dados geográficos, com foco no domínio de gerenciamento de desastres. Neste trabalho, ontologias de aplicação são criadas para descrever o modelo de dados usado por cada fonte de dados registrada no servidor. Para garantir a interoperabilidade de dados vindos de diferentes fontes, estas ontologias são definidas a partir de conceitos definidos em uma ontologia principal usada pelo servidor, que serve como um vocabulário compartilhado por todas as aplicações.

Depois que as ontologias de aplicação são definidas, a anotação temática de serviços espaciais é realizada através de descritores chamados de CSDs. Estes descritores são documentos XML baseados no padrão Dublin Core, e possuem um conjunto de metadados que descrevem informações a respeito da fonte de dados que oferece o serviço e os detalhes de acesso ao mesmo. As informações para o preenchimento do CSD são passadas pelo provedor do serviço no momento em que o mesmo é cadastrado no servidor. Neste momento, cada descritor é associado a uma ou mais classes da ontologia de aplicação usada pelo provedor que oferece o serviço que está sendo cadastrado. Estas classes descrevem o significado das informações que são oferecidas como saída no serviço. Finalizado o cadastro, o descritor é armazenado na base de dados do servidor.

No processo de recuperação de dados, usuários podem usar os conceitos definidos na ontologia para pesquisar por serviços que ofereçam o tipo de informação representada por este conceito. Além disto, o usuário pode criar seu próprio conceito de busca aplicando restrições aos conceitos definidos na ontologia compartilhada. Depois que o conceito de busca é definido pelo usuário, a aplicação analisa todos os CSDs cadastrados na base de dados e recupera todos os descritores que são associados a conceitos que são subsumidos pelo conceito de busca definido na consulta.

Embora o uso de ontologias melhore a recuperação de informações através de consultas com melhor cobertura, a falta de uma medida de similaridade faz com que todos os serviços recuperados sejam julgados com a mesma relevância para o usuário, o que pode prejudicar a localização de informações em consultas que recuperam uma grande quantidade de resultados. Outra desvantagem é que os CSDs anotam apenas as saídas de cada serviço, o que dificulta a possibilidade de descoberta serviços para criar uma composição.

3.1.1 O trabalho de Klien et al.

Klien et al. (2006) desenvolveram uma solução para a recuperação de serviços de processamento de dados geográficos, com foco no domínio de gerenciamento de desastres. Neste trabalho, ontologias de aplicação são criadas para descrever o modelo de dados usado por cada fonte de dados registrada no servidor. Para garantir a interoperabilidade de dados vindos de diferentes fontes, estas ontologias são definidas a partir de conceitos definidos em uma ontologia principal usada pelo servidor, que serve como um vocabulário compartilhado por todas as aplicações.

Depois que as ontologias de aplicação são definidas, a anotação temática de serviços espaciais é realizada através de descritores chamados de CSDs. Estes descritores são documentos XML baseados no padrão Dublin Core, e possuem um conjunto de metadados que descrevem informações a respeito da fonte de dados que oferece o serviço e os detalhes de acesso ao mesmo. As informações para o preenchimento do CSD são passadas pelo provedor do serviço no momento em que o mesmo é cadastrado no servidor. Neste momento, cada descritor é associado a uma ou mais classes da ontologia de aplicação usada pelo provedor que oferece o serviço que está sendo cadastrado. Estas classes descrevem o significado das informações que são oferecidas como saída no serviço. Finalizado o cadastro, o descritor é armazenado na base de dados do servidor.

No processo de recuperação de dados, usuários podem usar os conceitos definidos na ontologia para pesquisar por serviços que ofereçam o tipo de informação representada por este conceito. Além disto, o usuário pode criar seu próprio conceito de busca aplicando restrições aos conceitos definidos na ontologia compartilhada. Depois que o conceito de busca é definido pelo usuário, a aplicação analisa todos os CSDs cadastrados na base de dados e recupera todos os descritores que são associados a conceitos que são subsumidos pelo conceito de busca definido na consulta.

Embora o uso de ontologias melhore a recuperação de informações através de consultas com melhor cobertura, a falta de uma medida de similaridade faz com que todos os serviços recuperados sejam julgados com a mesma relevância para o usuário, o que pode prejudicar a localização de informações em consultas que recuperam uma grande quantidade de resultados. Outra desvantagem é que os CSDs anotam apenas as saídas de cada serviço, o que dificulta a possibilidade de descoberta serviços para criar uma composição.

3.1.2 O trabalho de Stock et al.

Outra solução para a recuperação de serviços foi desenvolvida por Stock et al. (2010). No lugar do uso de ontologias para a descrição dos modelos de dados, o trabalho propõe a criação de um catálogo de *feature types*, que serve como uma referência onde clientes podem encontrar a descrição de todos os tipos de dados oferecidos por uma IDE. Neste catálogo, cada *feature type* é descrito através de um conjunto de atributos, que representam as propriedades que o caracterizam, e um conjunto de operações, que representam todas as funções que podem ser aplicadas ao mesmo. Relacionamentos semânticos entre diferentes tipos de dados, como generalização e associação, também podem ser definidos. O modelo usado por este catálogo é criado através da extensão dos elementos do modelo de dados definido pelo catálogo ebRIM (OGC, 2005a).

A anotação dos recursos acontece através dos elementos que descrevem as operações oferecidas por cada *feature type*. Cada componente usado para a descrição de uma operação contém um conjunto de elementos de ligação, e cada elemento associa a operação a um serviço. Estes componentes representam o mapeamento entre a descrição abstrata de um serviço e as implementações disponíveis para o mesmo. Estes serviços, uma vez descobertos, podem ser diretamente acessados pelo usuário que realizou a consulta.

Para a recuperação de informações, um serviço de registro é oferecido. Este serviço pode ser acessado como um serviço *web*. O registro oferece um conjunto de funções que permitem ao usuário navegar sobre a hierarquia de *feature types* cadastrados no servidor. Dentre as operações permitidas estão a listagem dos tipos disponíveis, a recuperação de tipos que representam subclasses ou que estão associadas a um determinado *feature type*, entre outras. Outra função oferecida por este registro permite que o usuário recupere as informações de todos os serviços que implementam uma determinada operação escolhida pelo usuário.

A falta de ontologias para a descrição do tema dos recursos prejudica o processo de coleta de informações, pois dificulta a possibilidade de inferência de novos conhecimentos durante o processamento de consultas. Além disto, a solução oferece pouca flexibilidade, uma vez que o usuário não pode definir novos tipos de serviço quando o serviço oferecido não é descrito por qualquer operação definida no catálogo. Mais ainda, falta uma métrica para avaliar quais serviços recuperados são mais relevantes para o usuário.

3.1.3 O trabalho de Lemmens et al.

Lemmens et al. (2007) desenvolveram um *framework,* chamado SIFGEO, para resolver a interoperabilidade semântica de dados espaciais, bem como permitir a localização de serviços geográficos. A solução dos autores é composta por três ontologias. A primeira ontologia é chamada de ontologia de conceito de feições, e é usada para descrever domínios de aplicação. A segunda ontologia é chamada de ontologia de símbolo de feições, e é usada para descrever as características comuns a todos os *feature types*, com base no padrão ISO 19109 (ISO, 2005). A terceira ontologia é chamada de OPERA-R, e descreve um conjunto de operações que podem ser utilizadas para o processamento de feições geoespaciais.

A solução proposta pelos autores possui um serviço que armazena a descrição semântica dos serviços disponibilizados para o usuário. Os serviços são anotados e classificados de acordo com as ontologias oferecidas pelo *framework*. A anotação semântica dos serviços pode ser realizada com OWL-S ou através da anotação semântica do seu arquivo WSDL.

A resolução de uma consulta ocorre em três etapas. Na primeira, um protótipo chamado *GeoMatchMaker* realiza uma consulta sobre os serviços descritos na ontologia de serviços geográficos para verificar que tipos de serviços são necessários para resolver a consulta do usuário. Caso não haja um tipo de serviço capaz de resolver totalmente a consulta, o protótipo tenta encontrar composições de serviços que a satisfaçam. Caso a consulta seja resolvida, o resultado desta etapa é um serviço (ou uma composição de serviços) abstrata, sem estar associada a nenhuma implementação de serviço. De posse destas informações, a próxima etapa consiste em localizar implementações de serviço disponíveis para cada tipo de serviço selecionado na etapa anterior. Depois que estes serviços são localizados, um sistema de gerenciamento de *workflow* é responsável por controlar a execução do serviço ou da composição de serviços selecionada.

A solução proposta permite a recuperação de serviços e a descoberta de composições de serviços, levando em conta apenas a sua funcionalidade. Contudo, uma forma de encontrar serviços de dados geográficos se faz necessária. Tal procedimento ajudaria a encontrar serviços cujos dados poderiam ser usados para o processamento de cada um destes serviços. Para isto, precisa-se de um mecanismo que permita anotar a semântica de cada *feature type* oferecido pelo serviço.

3.1.4 O trabalho de Lutz

Outro trabalho proposto para a descoberta de serviços espaciais foi proposto por Lutz (2007). O objetivo deste trabalho é permitir a descoberta de serviços e composições de serviços que oferecem uma determinada informação solicitada pelo usuário.

Neste trabalho, os serviços de processamento de dados geográficos são anotados semanticamente através de uma assinatura semântica descrita através de lógica de primeira ordem. A assinatura contém informações a respeito dos parâmetros de entrada e saída do serviço, suas pré-condições e seus efeitos. Todas estas informações são descritas através de conceitos definidos em uma ontologia, criada a partir da série de padrões ISO 19100 para o domínio de dados geográficos. As consultas do usuário também são representadas através desta assinatura. O processo de descoberta de informações consiste em aplicar um *reasoning* sobre as descrições de serviços disponíveis a fim de encontrar descrições cuja assinatura é subsumida pela assinatura da consulta do usuário.

A principal vantagem deste trabalho é a proposição de uma solução para a descoberta de funcionalidades oferecidas por serviços ou composições de serviços. Contudo, não existem meios para encontrar serviços de dados espaciais que oferecem informações que poderiam ser usadas para a descoberta destas composições, uma vez que a assinatura não permite a anotação semântica de *feature types* oferecidos pelos serviços cadastrados.

3.1.5 O trabalho de Li et al.

Um importante trabalho que usa semântica para melhorar a descboerta de serviços geográficos oferecidos por uma IDE foi proposto por Li et al. (2011). O principal objetivo deste trabalho era oferecer uma infraestrutura para permitir a fácil localização de serviços que oferecem dados geográficos sobre a região ártica. Neste trabalho, uma ontologia de hidrologia é usada para melhorar a descrição da semântica dos serviços e o processo de descoberta de informações.

Para melhorar a descoberta de informações, o trabalho utiliza uma base de dados que contém informações sobre vários serviços de dados geográficos. Estes serviços podem ser coletados a partir de serviços de catálogos ou através de *crawlers*, que processam a URL das páginas disponíveis na Internet para identificar serviços de dados geográficos que não são cadastrados em nenhum catálogo. As informações contidas nos

metadados e no documento de funcionalidades do serviço são armazenadas e usadas durante o processamento de consultas.

Para descobrir os dados do seu interesse, o usuário acessa uma interface gráfica e seleciona, na ontologia usada pelo sistema, os conceitos e as propriedades que correspondem ao tipo de informação no qual está interessado. Os nomes destes conceitos e propriedades são passados para a ferramenta de busca da aplicação. Feito isto, a ferramenta realiza um *reasoning* sobre a ontologia para descobrir conceitos e propriedades que são relacionados ao conceito de busca selecionado. Uma vez localizados, os nomes destes conceitos e propriedades são usados para estender as palavras-chave contidas na consulta do usuário. Como resultado de uma consulta, a aplicação retorna todos os serviços (ou composição de serviços) que oferecem *feature types* que possuem em sua descrição os conceitos definidos na consulta estendida.

A principal vantagem deste trabalho consiste na utilização de ontologias e na verificação de informação em nível de *feature types*, o que melhora a qualidade das consultas. Entretanto, não são oferecidos meios para avaliar a relevância de cada resultado recuperado. Além disto, restrições espaciais e temporais são resolvidas apenas com base nas informações contidas nos metadados do serviço, o que reduz a qualidade deste tipo de consulta.

3.1.6 O trabalho de Chen et al.

Chen et al. (2011) desenvolveram uma abordagem que usa ontologias para melhorar a descoberta de serviços e *feature types* geográficos. Para alcançar este objetivo, foi implementado um catálogo que utiliza uma ontologia para descrever a semântica dos serviços WMS, WFS e WCS. Esta ontologia também armazena informações sobre cada *feature type* oferecido por estes serviços.

Para povoar o seu catálogo, a solução proposta utiliza um *crawler* para coletar informações sobre serviços disponíveis na Internet. Para cada serviço coletado, uma requisição *GetCapabilities* é executada para a obtenção do seu documento de funcionalidades. Uma vez recuperado, este documento é processado e as informações do serviço e de seus *feature types* são automaticamente convertidas para OWL e persistidas na ontologia usada pelo catálogo. Tal ontologia é criada através da extensão de OWL-S, e o mapeamento entre a descrição fornecida pelo documento de funcionalidades e a ontologia usada pelo catálogo é feito através de XSLT.

Durante o processo de resolução de consultas, o usuário pode especificar três tipos de restrição: o tipo de serviço requisitado, o título do *feature type* e a região geográfica de interesse. Ao receber a requisição, o serviço de catálogo executa um *reasoning* em sua ontologia para identificar todos os *feature types* que satisfazem os critérios da consulta. Ao fim da requisição, o catálogo recupera o perfil de todos os serviços que possuem os *feature types* selecionados durante a consulta. As informações destes serviços são então retornadas para o usuário.

A grande vantagem deste trabalho é que a utilização de ontologias, aliada ao armazenamento de informações em nível de *feature types*, permite melhorar a qualidade das consultas. Entretanto, uma importante limitação desta solução é que a ontologia proposta é muito geral e não fornece meios para identificar o domínio de aplicação de cada *feature type*. Ademais, a solução proposta não oferece *ranking* e não aborda a resolução de consultas com restrições temporais.

3.2 Recuperação de informação em nível de *feature types*

O problema da descoberta de *feature types* consiste em localizar, para o usuário, as camadas que oferecem a informação desejada pelo mesmo. A vantagem deste tipo de descoberta em relação à descoberta de serviços de dados geográficos é que logo após a consulta o usuário já sabe diretamente quais camadas oferecem os dados geográficos de seu interesse.

3.2.1 O trabalho de Bernard et al.

Bernard et al. (2006) desenvolveram uma abordagem baseada em ontologias para resolver o problema de recuperação e integração de dados geográficos. O principal objetivo deste trabalho é facilitar a localização e a integração de dados geográficos oferecidos por diferentes fontes de informação. Este trabalho está sendo aplicado à recuperação de dados de hidrografia.

Para implementar a sua solução, o trabalho propõe a utilização de dois tipos de ontologias. O primeiro tipo é usado para descrever a semântica dos *feature types* oferecidos por serviços WFS. Cada serviço WFS oferecido pela infraestrutura possui uma ou mais ontologias deste tipo. Tais ontologias são geradas a partir da análise dos esquemas XML obtidos para os *feature types* oferecidos pelo serviço. Tais esquemas são obtidos através da invocação da operação *DescribeFeatureType* oferecida pelo serviço WFS. O segundo tipo de ontologia é usado para descrever um vocabulário

compartilhado por todas as fontes de dados. Neste caso, para cada ontologia usada por um serviço, uma ontologia de mapeamento deve ser criada para descrever a correspondência entre os seus conceitos e os conceitos definidos na ontologia compartilhada.

Para localizar os dados do seu interesse, o cliente pode acessar o serviço de catálogo da IDE e formular consultas com base nos conceitos definidos na ontologia compartilhada. Nesta consulta, o cliente pode selecionar um dos conceitos definidos na ontologia ou criar um novo conceito com base nos conceitos oferecidos. Feito isto, a ferramenta de busca aplica um *reasoning* sobre a ontologia compartilhada para identificar todos os conceitos que são subsumidos pelo conceito definido na consulta do usuário. Depois, a ferramenta de busca recupera todos os serviços que oferecem *feature types* associados a estes conceitos.

A vantagem deste trabalho é a utilização de ontologias para melhorar a cobertura das consultas. Entretanto, a utilização de múltiplas ontologias para cada serviço pode requerer uma grande quantidade de mapeamentos durante a realização de uma consulta, o que prejudica a sua escalabilidade. Outras limitações importantes ocorrem porque a solução proposta não aborda a resolução de consultas com restrições geográficas e temporais, e tampouco oferece uma medida de *ranking* para avaliar a relevância dos resultados recuperados.

3.2.2 O trabalho de Lutz e Klien

Um trabalho proposto que usa ontologias para melhorar a localização de dados geográficos oferecidos por IDEs foi desenvolvido por Lutz e Klien (2006). A abordagem proposta pelos autores usa dois tipos de ontologias para descrever a semântica dos domínios de aplicação oferecidos pela infraestrutura. O primeiro tipo é chamado de ontologia de domínio, e corresponde a um vocabulário comum que é compartilhado por todas as fontes de dados da infraestrutura. O segundo tipo é chamado de ontologia de aplicação e descreve a visão do modelo usado por cada fonte de dados. Visando garantir a interoperabilidade entre estas ontologias, os conceitos definidos nas ontologias de aplicação são criados a partir dos conceitos definidos na ontologia de domínio.

A anotação semântica dos recursos oferecidos por cada fonte de dados é feita através de mapeamentos de registro (BOWERS; LUDÄSCHER, 2004). Neste tipo de anotação, cada *feature type* oferecido pela infraestrutura e suas respectivas propriedades

são associados, através de regras, a elementos definidos na ontologia de aplicação usada pela fonte de dados do seu provedor. Os mapeamentos de registro de cada *feature type* são armazenados em uma base de dados e usados para a resolução de consultas. Durante o processo de recuperação de informações, o usuário especifica o conceito de busca desejado, e o sistema recupera todos os *feature types* associados a conceitos que são subsumidos pelo conceito de busca definido na consulta.

A abordagem proposta pelos autores melhora a localização de dados geográficos, uma vez que o uso de ontologias para a realização deste processo melhora a qualidade das consultas. Contudo, como nenhuma medida de similaridade é oferecida, não há como distinguir a relevância de cada resultado recuperado para a consulta do usuário. Da mesma forma, a abordagem não permite a recuperação de recursos que satisfazem parcialmente os critérios de seleção definidos na consulta. Além disso, o trabalho não permite a recuperação de informação em nível de serviços e de feições.

3.2.3 O trabalho de Zhang et al.

Outra solução para a recuperação de *feature types* geográficos foi proposta por Zhang et al. (2010). Neste trabalho, uma ontologia espacial é usada para armazenar as informações referentes aos *feature types* oferecidos por serviços WFS registrados na IDE. Esta ontologia também descreve os modelos de dados referentes ao domínio de aplicação destes recursos. No trabalho em questão, é definida uma ontologia do domínio de transportes.

A anotação semântica dos *feature types* é realizada através da ontologia oferecida pela aplicação. Todos os *feature types* oferecidos pelo servidor são armazenados como instâncias de classes desta ontologia. Para cada *feature type*, são guardados o seu nome, o seu conjunto de propriedades, o seu tipo de geometria (ponto, linha, polígono) e o *bounding-box* referente à sua extensão geográfica. Além disto, cada *feature type* é associado a uma classe definida na ontologia do sistema. Esta classe indica qual conceito do domínio da aplicação representa a informação oferecida pelo *feature type*. Todo o processo de anotação semântica é realizado de forma automática pelo servidor, com base no nome do *feature type*.

Durante uma consulta, o usuário define as seguintes informações: o tipo de informação desejado (através da escolha de um dos conceitos da ontologia oferecida), as propriedades de seu interesse, o tipo de geometria desejado e o *bounding-box*. O processo de localização de informações consiste em recuperar todos os tipos de feição

cujas características são equivalentes ou são subsumidas pelas restrições definidas na consulta do usuário. A aplicação também permite a recuperação de recursos que casam parcialmente com a requisição do usuário. Para isto, uma medida de similaridade é usada para avaliar a relevância de cada recurso para a consulta do usuário. Tal relevância é calculada através da soma das relevâncias de cada característica considerada na consulta do usuário, na qual cada característica possui um peso e a soma dos pesos de todas as características é igual a 1. A similaridade entre a classe da consulta do usuário e a classe do *feature type* é calculada através do número de subclasses existentes entre as mesmas. A similaridade entre os conjuntos de propriedades e entre as geometrias é calculada de forma semelhante à utilizada para avaliar a similaridade entre os conceitos. Por fim, a similaridade entre os *bounding-boxes* é calculada através do tamanho da área de sobreposição entre os mesmos.

A grande vantagem deste trabalho é a possibilidade de recuperar recursos que satisfazem parcialmente as restrições da consulta, além da proposição de uma métrica para avaliar o *ranking* de cada *feature type* recuperado. Contudo, a solução proposta não considera as características temporais dos recursos para a realização de consultas. Além disso, a sua abordagem para avaliar a similaridade semântica apenas considera o relacionamento de generalização entre os componentes, não considerando outros tipos de relacionamento existentes.

3.2.4 O trabalho de Janowicz et al.

Outra solução baseada em similaridade para a recuperação de *feature types* em IDEs foi proposta por Janowicz et al. (2008). Neste trabalho, a similaridade entre conceitos definidos em ontologias é calculada com base em um *framework* chamado SIM-DL (JANOWICZ, 2006), que avalia a similaridade de conceitos definidos em *AllNR*, uma variação da linguagem *Description Logic*. A abordagem proposta pelos autores considera também que a infraestrutura oferece uma série de ontologias para descrever seus domínios de aplicação, e que cada *feature type* oferecido pela IDE é associado a um conceito definido em uma destas ontologias.

Durante uma consulta, o usuário especifica um conceito de busca e uma região geográfica de interesse. Com base nestas restrições, o sistema recupera todos os *feature types* que são associados a este conceito de busca. Contudo, o sistema também sugere ao usuário conceitos que não são equivalentes ao conceito de busca, mas que apresentam

grande similaridade com o mesmo. O usuário pode então selecionar e recuperar todos os *feature types* referentes a cada um dos conceitos sugeridos pelo sistema.

A grande vantagem deste trabalho é o desenvolvimento de uma medida de similaridade que permite avaliar o quanto cada conceito definido na ontologia usada pela infraestrutura é similar ao conceito definido na consulta do usuário. Além disto, *feature types* associados a conceitos relacionados ao conceito de busca também podem ser recuperados. Contudo, o trabalho é voltado apenas para uma linguagem de ontologia especifica. Mais ainda, as dimensões espaço e tempo não são consideradas durante o processo de localização de informações.

3.2.5 O trabalho de Wiegand e Garcia

Wiegand e Garcia (2007) desenvolveram uma abordagem que usa tarefas para a recuperação de informações. O modelo semântico usado para a implementação deste trabalho é baseado em quatro tipos de ontologias. A primeira delas é uma ontologia de tarefas, que representa uma taxonomia de serviços oferecidos pela infraestrutura, como o gerenciamento de emergência, o planejamento de uso da terra, entre outros. A segunda é uma ontologia de fontes de dados, que descreve os domínios de aplicação disponíveis em uma infraestrutura, como agricultura, transporte, relevo, entre outros. O terceiro tipo é uma ontologia de metadados, que representa uma taxonomia que contém os metadados descritivos de uma fonte de dados, de acordo com o padrão de metadados definido pelo FGDC (*Federal Geographic Data Committee*). O último tipo corresponde a uma ontologia de lugar, que descreve as características referentes à localização dos recursos cadastrados no servidor. Todas estas ontologias são conectadas entre si. Por exemplo, cada instância de uma tarefa é composta por um conjunto de classes da ontologia de fonte de dados. Estas classes representam os *feature types* que são necessários para a realização da tarefa. Para cada um destes tipos, um conjunto de restrições pode ser definido através de regras. Por sua vez, as instâncias das classes da ontologia de fontes de dados são associadas a uma descrição de metadados e a uma determinada região geográfica.

A principal característica deste trabalho é que as tarefas são definidas em tempo de projeto. Desta forma, um *expert* de domínio pode especificar um conjunto de tarefas oferecidas pela fonte de dados. Para cada tarefa, são definidos os *feature types* que precisam estar disponíveis para a sua execução. Esta característica permite que, em situações de emergência, como, por exemplo, durante a ocorrência de um desastre, os

responsáveis por tomar as ações necessárias para a sua resolução tenham acesso rápido às fontes de dados que podem ser usadas, sem a necessidade de ter que descobrir e avaliar estas informações em tempo de execução.

Durante o processo de localização de informações, o usuário seleciona uma das tarefas oferecidas pela ontologia de tarefas da infraestrutura e define a região geográfica de seu interesse. Depois que esta seleção é feita, um *reasoning* é realizado sobre as instâncias cadastradas em cada ontologia, a fim de recuperar os conjuntos de dados que oferecem as informações necessárias para a realização da tarefa solicitada e que estão associados à região geográfica delimitada na consulta. Tal recuperação é feita através do relacionamento semântico de subsunção.

A desvantagem deste trabalho é que a localização de informações só pode ser realizada a partir de tarefas. Mais ainda, faltam mecanismos que permitam avaliar, quando dois ou mais serviços disponíveis oferecem a informação necessária para a realização da tarefa, qual pode ter mais relevância para o usuário.

3.3 Recuperação de informação em nível de feições

O problema da recuperação em nível de feições consiste em localizar *features* espaciais que satisfazem os critérios de busca definidos pelo usuário. Este tipo de recuperação é utilizado para *feature types* oferecidos por serviços WFS. O uso de semântica para resolver este tipo de problema requer o uso de mapeamentos entre a linguagem usada para a definição da ontologia e a linguagem GML, usada para o acesso a este tipo de informação no serviço.

Uma forma de implementar este tipo de recuperação consiste em definir as restrições do usuário sobre os conceitos da ontologia e mapear estas restrições para GML através de um filtro OGC, que é aplicado para recuperar, em GML, apenas as feições de interesse do usuário. Outra solução possível consiste em recuperar todas as feições de um determinado *feature type* em formato GML e convertê-las para instâncias da linguagem usada pela ontologia. Depois, um *reasoning* pode ser aplicado para verificar quais feições satisfazem as restrições definidas pelo usuário.

3.3.1 O trabalho de Lutz e Kolas

Uma abordagem baseada em ontologias e regras para a recuperação de informação em nível de feições foi proposta por Lutz e Kolas (2007). Neste trabalho,

regras são utilizadas para descrever os conceitos que compõem a ontologia do domínio de aplicação da infraestrutura. Esta ontologia é usada como base para a recuperação de informações. A anotação das fontes de dados acontece através de regras de mapeamento. Cada fonte de dados que compõe a IDE deve descrever um conjunto de regras de mapeamento, que descrevem como os seus *feature types* podem ser mapeados para a ontologia usada pela infraestrutura.

Para recuperar informações, o usuário pode selecionar, na ontologia da infraestrutura, o conceito no qual está interessado e definir uma ou mais restrições que devem ser satisfeitas por suas feições. Feito isto, a ferramenta de busca analisa a descrição do conceito selecionado e expande a consulta para outros conceitos que são semanticamente relacionados ao mesmo. Depois, a aplicação consulta todas as fontes de dados cadastradas e recupera todos os *feature types* que podem ser relevantes para a consulta do usuário. Quando um *feature type* é considerado relevante, suas feições são mapeadas de GML para instâncias de conceitos da ontologia da IDE. As instâncias recuperadas em cada fonte de dados são carregadas para um servidor central, no qual formam uma base de conhecimento que é usada como suporte para a resolução da consulta. Depois que esta base é montada, um *reasoning* é realizado para determinar as instâncias que satisfazem os critérios de busca definidos na consulta.

A abordagem proposta por este trabalho melhora a recuperação em nível de feições através do uso de ontologias. Contudo, o fato de que muitos dados precisam ser convertidos para instâncias da ontologia e carregados para uma base de dados centralizada durante a realização de cada consulta compromete a sua escalabilidade. Ademais, o trabalho não aborda restrições espaciais e temporais, e não oferece uma medida de *ranking* para avaliar os resultados recuperados.

3.3.2 O trabalho de Batcheller e Reitsma

Batcheller e Reitsma (2010) desenvolveram um trabalho que usa ontologias para melhorar a recuperação de dados geográficos. A aplicação é voltada para o aproveitamento de bases de dados georeferenciadas mantidas por sistemas legados. A ideia do trabalho consiste em converter os dados existentes para triplas RDF sob demanda, à medida em que os mesmos vão sendo requisitados pelos clientes. Para isto, uma ontologia é usada para descrever a semântica do domínio de aplicação referente aos dados existentes.

A anotação semântica dos dados é realizada através de um arquivo de mapeamento. O objetivo deste arquivo é especificar como deve ocorrer a tradução entre as classes (e suas respectivas propriedades) definidas na ontologia e os *feature types* (e seus respectivos atributos) existentes na base de dados. Estes mapeamentos são especificados através da linguagem de mapeamento da plataforma D2RQ, que é usada para fazer a tradução da linguagem usada pela ontologia para a linguagem usada na base de dados.

Durante uma consulta, o usuário seleciona, na ontologia, o conceito referente ao tipo de informação que ele deseja. O usuário pode também especificar um conjunto de restrições sobre as propriedades do conceito escolhido. Definidas as restrições, uma consulta na linguagem SPARQL é usada para representar a requisição. Quando a aplicação recebe a consulta, os arquivos de mapeamento são verificados para avaliar quais as tabelas da base de dados que contém as informações relativas aos conceitos definidos na consulta do usuário. As tuplas destas tabelas são então recuperadas e convertidas para triplas RDF. Feito isto, um *reasoning* é usado para encontrar as instâncias que satisfazem os critérios de busca definidos na consulta do usuário, que são recuperadas e apresentadas para o mesmo.

A abordagem usada pelos autores melhora a localização de dados geográficos oferecidos por sistemas legados. Contudo, o fato de que os dados precisam ser pré-carregados para uma base central para o processo de *reasoning* prejudica a sua escalabilidade para que a solução seja implantada em um ambiente como IDEs, nas quais a recuperação de instâncias pode envolver a utilização de informações vindas de várias bases de dados heterogêneas.

3.4 Abordagens genéricas

Alguns trabalhos propostos para a recuperação de informações não focam na localização de um tipo de recurso especifico. Geralmente, estes trabalhos propõem soluções que são capazes de recuperar qualquer recurso cadastrado no catálogo da infraestrutura. O uso de semântica para a solução deste problema é feito através da associação do registro de metadados de cada recurso a conceitos definidos em ontologias.

3.4.1 O trabalho de Athanasis et al.

Athanasis et al. (2009) propõem uma solução baseada em semântica para melhorar a recuperação de recursos oferecidos por portais geográficos. Esta solução está sendo aplicada a um portal para a localização de dados relativos a desastres naturais. A abordagem usada por este trabalho usa três ontologias, que são representadas na forma de esquemas RDF. O primeiro esquema é usado para classificar o formato em que va informação é oferecida, como, por exemplo, imagem, *shapefile*, documento, etc. O segundo esquema é usado para modelar domínios de aplicação e classificar o tipo de informação oferecida pelo recurso, como desastre, incêndio, e inundação. O terceiro esquema é uma taxonomia contendo os elementos do padrão ISO 19115.

Neste trabalho, quando um provedor de informação cadastra seus recursos no portal, ele precisa associar cada recurso cadastrado a classes definidas nos dois primeiros esquemas, além de fornecer as informações referentes ao terceiro esquema. Todas as informações passadas durante o registro são convertidas para RDF e armazenadas em uma base de dados.

Durante o processo de recuperação de informações, o usuário pode navegar, através de *browsing*, sobre as classes de cada um destes esquemas e escolher aquelas referentes ao tipo de informação na qual ele está interessado. Para cada classe selecionada, o usuário pode definir restrições que devem ser satisfeitas pelas suas instâncias. Depois que as classes e suas respectivas restrições foram definidas, a consulta do usuário é convertida para RQL e comparada com as descrições de recursos cadastradas na base de dados. O sistema recupera os metadados de todos os recursos cuja descrição satisfaz a todas as restrições definidas pela consulta do usuário.

A desvantagem desta solução é que ela não permite distinguir o grau de relevância de cada recurso recuperado, e não oferece a flexibilidade da recuperação de informações com casamento parcial.

3.4.2 O trabalho de Smits e Friis-Christensen

Smits e Friis-Christensen (2007) desenvolveram um trabalho que usa semântica para realizar a recuperação de recursos em uma IDE européia. Para realizar esta tarefa, o trabalho usa um grafo conceitual, que é criado a partir de um *thesaurus* multilíngue. Este grafo contém o conjunto de conceitos usado para modelar os domínios de aplicação

oferecidos pela IDE. Mais ainda, relacionamentos semânticos entre estes conceitos são representados através de arcos neste grafo conceitual.

Para realizar a anotação semântica, cada recurso cadastrado na IDE possui um conjunto de descritores, baseados no padrão de metadados Dublin Core, que descrevem suas características (nome, título, abstract, palavras-chave e extensão geográfica) através de linguagem natural. A aplicação proposta pelo trabalho processa as informações textuais contidas nestes descritores para verificar o tipo de informação que é oferecida pelo recurso. O objetivo desta etapa é encontrar os conceitos do grafo conceitual que representam o conteúdo da informação oferecida pelo recurso. Depois que estes conceitos são encontrados, eles são associados aos descritores do recurso. Estas associações são armazenadas numa base de dados e usadas durante o processamento de consultas. Durante o processo de busca, o usuário pode navegar através dos conceitos definidos neste grafo, selecionar um conceito de busca desejado, e recuperar todos os recursos que possuem algum descritor associado ao mesmo.

A desvantagem desta solução é que não são oferecidos meios para distinguir o grau de relevância de cada recurso recuperado, e não é oferecida a flexibilidade da recuperação de informações com casamento parcial. Além disto, a recuperação em nível de feições não é oferecida.

3.4.3 O trabalho de Macário et al.

Macário et al. (2009) desenvolveram um trabalho que usa anotações semânticas para a recuperação de informações referentes a interpretações de dados geoespaciais. Neste trabalho, ontologias de domínio são usadas para anotar as informações relativas a recursos disponibilizados pelas fontes de dados que se registram na aplicação.

A anotação semântica das informações é realizada através de um conjunto de *workflows* oferecidos pela aplicação, que são criados por *experts* de domínio. Cada *workflow* é voltado para a anotação semântica de um determinado tipo de informação. Quando um novo recurso é cadastrado, o usuário seleciona e executa, dentre os *workflows* disponíveis, aquele que é voltado para o tipo de recurso que está sendo cadastrado. Durante o processo de anotação, o provedor que está registrando seus recursos oferece todas as informações necessárias para a realização desta tarefa. Este processo gera um conjunto de anotações, que são armazenadas em triplas RDF contendo a identificação do recurso, o *label* do elemento de metadados que está sendo anotado e o seu correspondente valor. Durante este processo, a aplicação também pode solicitar ao

usuário que o mesmo cadastre informações referentes à proveniência dos dados, região geográfica e como os dados obtidos podem ser interpretados. Por fim, as anotações geradas ao fim desta etapa são associadas a conceitos definidos em ontologias através de um serviço chamado Aondê (DALTIO; MEDEIROS, 2008).

Uma vez armazenados, os dados referentes às anotações podem ser recuperados através de linguagens de consulta baseadas em XML, como XPath ou XQuery. Contudo, a solução proposta não oferece meios para mensurar a similaridade de cada recurso oferecido. Outra limitação é que a estratégia usada para a anotação de informações dificulta a recuperação de informações em nível de *feature types* e em nível de feições.

3.5 Considerações finais

Este capítulo descreveu alguns dos principais trabalhos que compõem o estado da arte da pesquisa relativa à recuperação de dados e serviços geográficos. O capítulo mostrou que os trabalhos existentes variam de acordo com o tipo de recurso descoberto, podendo ser voltados para a localização de serviços de dados geográficos, serviços de processamento de dados espaciais, *feature types*, feições ou podem usar abordagens mais genéricas.

A discussão destes trabalhos também mostrou que nenhuma das soluções atuais resolve completamente o problema proposto, o que mostra a relevância do trabalho de pesquisa descrito nesta tese. Como um resumo, a Tabela 3.1 mostra uma comparação dos trabalhos discutidos, com relação a algumas das questões envolvidas na resolução do problema de pesquisa proposto. Os critérios usados para a avaliação foram selecionados a partir das principais questões que envolvem a recuperação de dados geográficos, bem como a partir dos requisitos levantados para o arcabouço proposto pela tese. As colunas desta tabela possuem os seguintes significados:

- **Trabalho:** representa o trabalho que está sendo considerado para a comparação. Todos os trabalhos mostrados na tabela foram apresentados ao longo deste capítulo. A última linha representa o SESDI, que é o trabalho proposto nesta tese;
- **C_1 (Uso de ontologias):** distingue se o trabalho usa ontologias para realizar a anotação e a recuperação de recursos;
- **C_2 (Recuperação de serviços):** identifica se o trabalho oferece suporte para a recuperação de informações em nível de serviços;

- **C_3 (Recuperação em nível de *feature types*):** identifica se o trabalho oferece suporte para a recuperação de informações em nível de *feature types* geográficos, que correspondem às camadas oferecidas por um serviço WMS ou WFS;
- **C_4 (Recuperação em nível de feições):** identifica se o trabalho oferece suporte para a recuperação de informações em nível de feições, que correspondem aos dados geográficos oferecidos por uma camada;
- **C_5 (Restrições temporais):** identifica se o trabalho oferece suporte para a resolução de consultas com restrições temporais;
- **C_6 (*Ranking*):** identifica se o trabalho oferece uma medida que permita avaliar a similaridade entre cada recurso oferecido e a consulta do usuário, permitindo que os resultados recuperados possam ser organizados em um *ranking* antes de serem apresentados ao usuário;
- **C_7 (Anotação Automática):** identifica se o trabalho consegue associar automaticamente os recursos oferecidos pela IDE ou portal geográfico a conceitos definidos em ontologias.

Tabela 3.1: Comparação entre os trabalhos discutidos

Trabalho	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Klien et al.	sim	sim	não	não	não	não	não
Stock et. al.	não	sim	não	não	não	não	não
Wiegand e Garcia	sim	não	não	não	sim	não	não
Lemmens et al.	sim	sim	não	não	não	não	não
Lutz	sim	sim	não	não	não	não	não
Lutz e Klien	sim	sim	não	não	não	não	não
Zhang et al.	sim	sim	não	não	não	sim	sim
Janowicz et al.	sim	não	não	não	não	sim	não
Lutz e Kolas	sim	não	sim	sim	não	não	não
Batcheller e Reitsma	sim	não	sim	sim	não	não	não
Athanasis et. al	sim	sim	sim	não	sim	não	sim
Smits e FriisChristensen	sim	sim	não	não	não	não	não
Macário et al.	sim	sim	não	não	não	não	não
Bernard et al.	sim	sim	não	não	não	não	não
Li et al.	sim	sim	não	não	não	não	não
Chen et al.	sim	sim	não	não	não	não	não
SESDI	sim	sim	sim	não	sim	sim	sim

Capítulo 4 – SESDI: Especificação

Este capítulo apresenta o arcabouço SESDI (*Semantic-Enabled Spatial Data Infrastructures*), que é a solução proposta por esta tese para resolver o problema da localização de dados geográficos oferecidos por uma infraestrutura de dados espaciais. Esta solução é um arcabouço que permite que os dados oferecidos pelo serviço de catálogo de uma IDE possam ser mais facilmente localizados pelos seus clientes. Especificamente, o capítulo trata das questões relacionadas à especificação do SESDI, discutindo os requisitos que nortearam o seu desenvolvimento, o modelo proposto para a recuperação de dados geográficos e o projeto arquitetural.

4.1 Levantamento de requisitos

A primeira etapa da especificação consistiu em definir os requisitos funcionais que deveriam ser satisfeitos pelo SESDI. Os requisitos levantados foram: a criação de uma nova base de dados, a localização de *feature types* geográficos, o uso de ontologias, a localização de informação baseada em *ranking,* o desenvolvimento de métricas para avaliar o *ranking* espacial, semântico, temporal e global, e a automatização do processo de coleta de informações.

4.1.1 Criação de uma nova base de dados sobre os serviços

Conforme discutido no Capítulo 1, muitas das limitações dos catálogos atuais acontecem porque as informações contidas nos seus registros de metadados fornecem poucos detalhes sobre as características espaciais, temáticas e temporais dos *feature types* oferecidos pelos serviços de dados geográficos. Para superar estas limitações, foi definido que uma nova base de dados teria que ser desenvolvida, visando o armazenamento de informações adicionais sobre os serviços cadastrados na infraestrutura. Esta nova base de dados deveria conter informações acerca da região geográfica, do tema e da extensão temporal de cada camada oferecida pelos serviços.

Ainda sobre este requisito, foi definido que a nova base de dados deveria conter apenas as informações necessárias para melhorar a resolução de consultas. Desta forma, o serviço de catálogo da IDE ainda seria responsável por manter as informações mais detalhadas dos serviços, incluindo as suas informações de qualidade, proveniência e as questões relativas aos direitos de acesso e de uso dos dados. Entretanto, para cada

serviço armazenado na nova base de dados, uma referência deveria ser armazenada para o seu registro de metadados, de forma a permitir que os clientes pudessem facilmente recuperar as informações adicionais sobre os serviços que oferecem cada camada recuperada em uma consulta.

4.1.2 Localização em nível de *feature types*

O arcabouço proposto deveria ser capaz de localizar *feature types* de interesse do usuário. Esta localização deveria acontecer através da análise das camadas oferecidas por cada serviço cadastrado na infraestrutura. O resultado deste tipo de consulta seria um conjunto contendo todas as camadas que satisfizessem os critérios de seleção estabelecidos na requisição do usuário. Para cada camada recuperada por uma consulta, deveriam ser mostradas também as informações sobre o serviço pelo qual a mesma é oferecida.

4.1.3 Uso de ontologias

Uma importante limitação dos serviços de catálogo atuais é a falta de meios formais para descrever o tema dos dados e serviços disponibilizados por seus provedores. Visando superar esta limitação, foi definido que ontologias deveriam ser usadas para descrever os domínios de aplicação de atuação da infraestrutura. O uso de ontologias, além de melhorar o processo de recuperação da informação, facilita a integração de informações oferecidas por diferentes fontes de dados.

4.1.4 Recuperação baseada em *ranking*

Muitas vezes, as consultas solicitadas pelos usuários podem gerar resultados que contêm uma grande quantidade de recursos. Neste tipo de situação, o usuário pode gastar uma grande quantidade de tempo para encontrar os dados desejados ou, em um caso pior, pode fazer com que estes dados nem mesmo sejam encontrados. Para minimizar os problemas causados por este tipo de situação, foi definido que todo o processo de recuperação de informações deveria ser baseado em *ranking*. Desta forma, uma métrica deveria ser desenvolvida para avaliar a relevância de cada *feature type* recuperado para a consulta do usuário. O objetivo deste requisito era fazer com que as camadas provavelmente mais relevantes para o cliente fossem apresentadas primeiro durante a exibição do resultado de uma consulta.

4.1.5 Anotação automática

Atualmente, quando um provedor de dados geográficos inclui o seu serviço no catálogo da IDE, precisa fornecer uma extensa lista de metadados sobre o mesmo. Desta forma, a solicitação de informações adicionais sobre a região geográfica, o tema e o período de tempo referente a cada uma de suas camadas implicaria em uma grande carga adicional de trabalho para o provedor, e diminuiria consideravelmente a aplicabilidade do arcabouço proposto. Visando evitar este tipo de limitação, foi definido que todas as informações adicionais sobre os serviços e suas camadas, necessárias para o modelo usado pelo SESDI, deveriam ser extraídas e identificadas de forma automática a partir das informações já disponibilizadas pelo provedor do serviço.

4.2 Um modelo baseado na recuperação da informação clássica

Uma característica importante da abordagem usada para a implementação do SESDI, é que os recursos usados, para realizar a recuperação de informações, são modelados de acordo com algumas ideias usadas nos modelos clássicos da recuperação da informação (do inglês *Information Retrieval, IR*). Destarte, algumas características dos modelos propostos para a recuperação de documentos foram adaptadas e utilizadas para melhorar a recuperação de informações no domínio de informações geográficas.

O modelo usado pelo SESDI considera que, da mesma forma que a *web* pode ser vista como uma coleção de documentos, uma IDE pode ser vista como um conjunto de serviços de dados geográficos. Além disto, na recuperação da informação clássica, cada documento é representado por um conjunto de termos, que podem ser usados para a sua indexação e recuperação. Analogamente, em uma IDE, cada serviço de dados geográficos pode ser visto como uma coleção de *feature types*. Assim, as características destas camadas podem ser usadas para a indexação e recuperação destes serviços. É importante ressaltar que o formato em que um *feature type* é representado depende do tipo de serviço que o oferece. Por exemplo, no caso de serviços WMS, cada camada é oferecida como uma imagem ou através da linguagem SVG. No caso de serviços WFS, cada camada é representada no formato GML, *shapefile* ou qualquer outro formato aceito pelo serviço. A comparação entre o modelo usado para a recuperação da informação clássica e o modelo proposto para melhorar a recuperação de informações geográficas em uma IDE é mostrada na Figura 4.1.

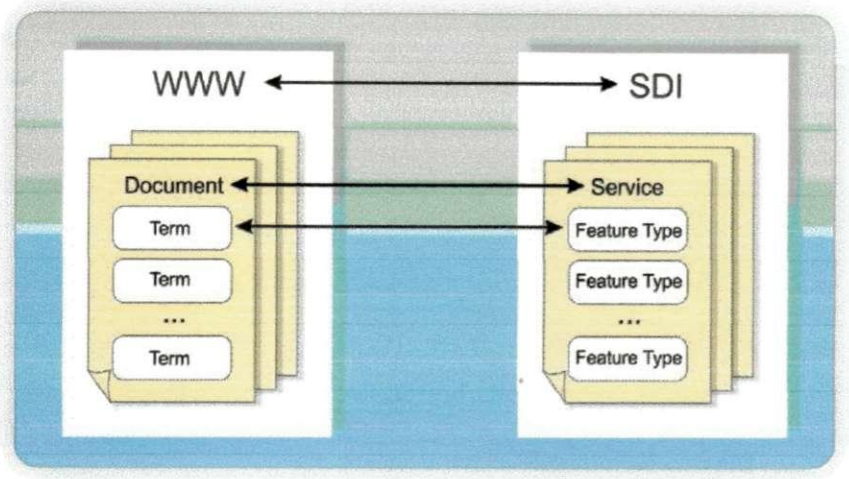

Figura 4.1: Comparação entre a recuperação clássica e a abordagem proposta

Apesar das semelhanças discutidas acima, também existem algumas diferenças entre a IR clássica e o domínio de dados geográficos. A principal delas é que, diferentemente de palavras-chave, que correspondem apenas a um texto simples, *feature types* geográficos são estruturas mais complexas e caracterizadas por três dimensões: espaço, tema e tempo. Desta forma, mecanismos mais complexos precisam ser desenvolvidos para avaliar se um determinado *feature type* satisfaz os critérios definidos na requisição do usuário. Como as consultas do usuário podem ter restrições sobre qualquer uma destas três dimensões, são necessários instrumentos para avaliar este casamento de acordo com cada uma destas dimensões.

Outra diferença importante é que na IR, os documentos julgados relevantes são sempre recuperados por completo. No domínio de dados geográficos, contudo, existem várias situações em que o usuário não está interessado em todo o serviço, mas em apenas uma ou mais camadas oferecidas pelo mesmo. Desta forma, são necessários meios para localizar, dentre todos os *feature types* oferecidos pelo serviço, apenas aqueles que são de interesse do usuário. Estas diferenças fizeram surgir a necessidade de algumas adaptações durante o desenvolvimento da solução proposta. Tais adaptações são explicadas com mais detalhes nas próximas seções.

Depois de analisar as semelhanças e diferenças entre a IR clássica e o domínio geoespacial, um modelo foi proposto para representar as informações oferecidas por uma IDE. Este modelo é descrito a seguir. É importante ressaltar que esta definição corresponde a uma adaptação do modelo proposto por Baeza-Yates e Ribeiro-Neto (1999), para a recuperação da informação, ao domínio de informações geográficas.

Definição 1: Um modelo de recuperação de informação geográfica para a localização de dados oferecidos por uma insfraestrutura de dados espaciais é uma quíntupla {S, T, Q, F, R (Qi, Tj)}. Segue a definição:

- S corresponde ao conjunto formado por todos os serviços de dados geográficos (WMS e WFS) descritos no serviço de catálogo da infraestrutura;
- T é a coleção composta por todos os *feature types* oferecidos pela IDE. Neste conjunto, cada *feature type* $Ti \in T$ e é oferecido por um serviço $Sj \in S$;
- Q é a representação de uma consulta do usuário;
- F é um *framework* que modela os serviços, os *feature types* e as consultas do usuário;
- R (Q_i, T_j) é uma função que determina o quanto um *feature type* T_j, pertencente ao conjunto T, é relevante para uma consulta Q_i formulada por um cliente. O valor desta relevância deve ser um número real entre 0 e 1. O valor 0 indica que o *feature type* não tem qualquer relevância para a consulta, enquanto o valor 1 indica que o recurso satisfaz totalmente todos os critérios de seleção definidos pela mesma.

Depois de definir o modelo para a recuperação de informação geográfica, foi criado um esquema conceitual para descrever as informações que seriam armazenadas na base de dados do SESDI e como estas informações estariam organizadas. O esquema desenvolvido, além de ser baseado no modelo proposto pela tese, satisfaz os requisitos funcionais especificados para o arcabouço.

Uma visão geral do esquema conceitual usado para a implementação da base de dados é mostrado na Figura 4.2. É importante ressaltar que a simplicidade do esquema deve-se ao fato de que a maior parte das informações acerca dos serviços continua sendo armazenada no serviço de catálogo da IDE. Assim, a base de dados desenvolvida para o SESDI armazena apenas as informações necessárias para melhorar a resolução das consultas, além de algumas informações preliminares (nome, título, descrição textual, palavras-chave) que podem ser usadas para o cliente fazer uma avaliação inicial dos *feature types* recuperados por uma consulta. Entretanto, o modelo armazena, para cada serviço, o identificador do seu registro de metadados. Assim, este registro pode ser

facilmente recuperado caso o cliente esteja interessado em informações mais detalhadas sobre o serviço, como, por exemplo, o processo de produção dos dados e as políticas de utilização dos seus dados.

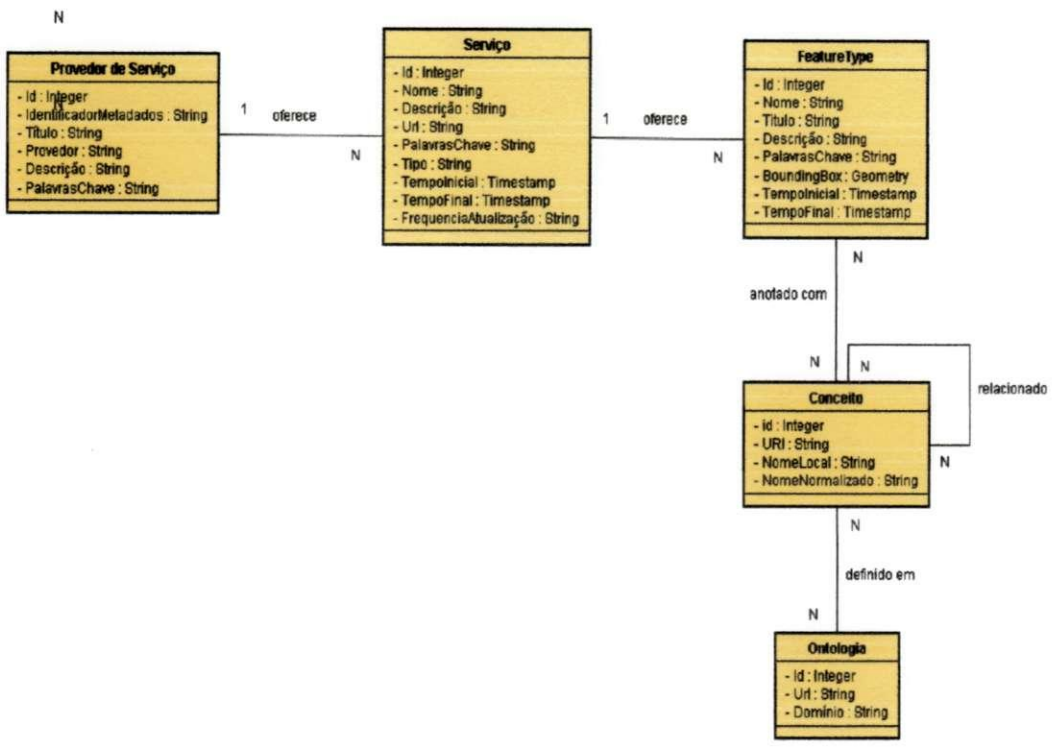

Figura 4.2: Esquema da base de dados

A análise do esquema conceitual permite perceber que a base de dados armazena informações sobre todas as ontologias usadas pelo arcabouço. Para cada ontologia, esta base também armazena informações sobre os seus conceitos. O motivo do armazenamento destas informações é a possibilidade de resolver consultas baseadas em ontologias sem a necessidade de realizar *reasoning* em tempo de execução, o que permite acelerar a resolução de consultas temáticas e aumentar a escalabilidade da solução. Mais detalhes sobre o uso de ontologias para a anotação temática de *feature types* e para a resolução de consultas são mostrados nas próximas seções.

Além das informações sobre cada ontologia, a base de dados usa a entidade *Provedor de Serviço* para armazenar informações sobre os provedores de dados geográficos. Para esta entidade, são armazenados atributos como a sua identificação única na base de dados, o identificador do seu registro no serviço de catálogo, o título do registro, o nome do seu provedor, a sua descrição e as suas palavras-chave.

A entidade *Serviço* representa os serviços de dados geográficos oferecidos pelos provedores cadastrados na IDE. As instâncias desta entidade formam o conjunto S do modelo proposto por esta tese. Para os serviços, são guardadas informações como a sua identificação única, o nome, a descrição textual, a URL de acesso, o tempo inicial, o tempo final, e a frequência de atualização. Estes três últimos atributos são usados para a resolução de consultas temporais, e serão explicados em mais detalhes no próximo capítulo.

As camadas oferecidas por cada serviço são representadas pela entidade *Feature Type*. Esta entidade desempenha um papel fundamental para o arcabouço, uma vez que as suas instâncias são utilizadas como base para a resolução das consultas. Uma característica importante desta entidade é o armazenamento de informações que descrevem as dimensões espacial, temática e temporal de cada *feature type*. A dimensão espacial é armazenada através do atributo *BoundingBox*, que descreve o menor retângulo que cobre a sua região geográfica. Este atributo armazena a geometria desta região. A dimensão temporal, por sua vez, é representada através dos atributos *TempoInicial* e *TempoFinal*, que descrevem o período de tempo referente aos dados oferecidos pelo mesmo. Por fim, a dimensão temática é descrita fazendo-se a associação do *feature type* a conceitos definidos nas ontologias usadas pelo SESDI. O esquema mostra que cada camada pode ser associada a um conjunto de conceitos, que descrevem o significado dos dados oferecidos pelo mesmo.

4.3 Projeto arquitetural

Outro importante artefato produzido durante a especificação do SESDI foi o projeto arquitetural. Esta arquitetura, que é mostrada na Figura 4.3, contempla os principais módulos e os seus relacionamentos com os componentes existentes da IDE.

O módulo de gerenciamento de ontologias é utilizado pelo administrador do SESDI. Este módulo oferece funções que permitem que novas ontologias sejam adicionadas a sua base de dados, além de permitir que as ontologias já existentes sejam modificadas ou excluídas. Sempre que uma nova ontologia é incluída, os seus conceitos passam a ser utilizados tanto para o processo de anotação temática dos *feature types* quanto para o processo de resolução de consultas.

O módulo de registro é responsável por identificar e coletar as informações sobre serviços e *feature types* oferecidos pela IDE. Para identificar estas informações, o módulo interage diretamente com o serviço de catálogo oferecido pela infraestrutura, no

qual procura por informações sobre novos serviços que foram adicionados ou que foram recentemente alterados. Depois de identificado, cada novo serviço é acessado para a recuperação do seu documento de funcionalidades. As informações obtidas a partir deste documento, juntamente com os dados obtidos a partir do catálogo, são usadas para identificar a extensão geográfica, temporal e para realizar a anotação temática dos *feature types* recuperados. Feito isto, todas as informações coletadas e geradas pelo módulo são armazenadas na base de dados do arcabouço.

Para realizar a resolução de consultas, o SESDI oferece uma interface gráfica, implementada com páginas *web* dinâmicas, que permite que usuários realizem consultas por dados geográficos de interesse. As páginas oferecidas permitem que os usuários definam as restrições espaciais, temáticas e temporais que devem ser satisfeitas pelos recursos. Definidos estes critérios, a consulta é enviada para a máquina de busca, que é módulo que contém os componentes usados para a resolução de consultas espaciais, temáticas, temporais e globais. A Figura 4.3 mostra que a máquina de busca também pode interagir com o serviço de catálogo. Tal interação acontece quando o usuário requisita informações mais detalhadas sobre um *feature type* ou serviço recuperado por uma consulta.

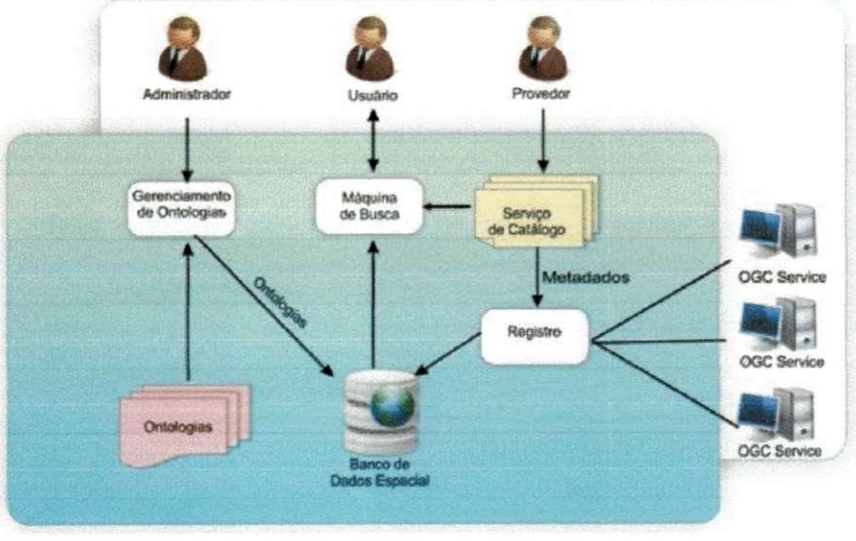

Figura 4.3: Projeto Arquitetural do SESDI

4.4 Considerações finais

Este capítulo apresentou o arcabouço SESDI, que representa a solução proposta por esta tese para melhorar a localização de dados geográficos oferecidos por uma

infraestrutura de dados espaciais. O capítulo discutiu a especificação do arcabouço proposto. Ao longo do seu desenvolvimento, foram apresentados os requisitos que nortearam o seu desenvolvimento, o modelo proposto para melhorar a recuperação de dados geográficos e o seu projeto arquitetural. O próximo capítulo descreve o processo de coleta de informações, destacando como as informações usadas pelo modelo usado pelo SESDI são obtidas a partir do serviço de catálogo da infraestrutura.

Capítulo 5 – SESDI: O processo de coleta de informações

Este capítulo descreve como as informações que compõem o esquema usado pelo SESDI são obtidas a partir do serviço de catálogo da infraestrutura. Ao longo do seu desenvolvimento, o capítulo discute a implementação da anotação espacial, temática e temporal das camadas oferecidas por cada serviço registrado no serviço de catálogo da IDE.

5.1 O processo de coleta de informações

O módulo de registro é responsável pela aquisição de informações geográficas junto ao serviço de catálogo da infraestrutura. O processo de coleta de informações é composto por quatro módulos:

- **gerenciamento:** este módulo gerencia o subsistema de coleta de informações, controlando e coordenando todo o processo e a interação entre os demais módulos;
- **aquisição de serviços:** este módulo é responsável por identificar novos serviços de dados geográficos adicionados ao serviço de catálogo da IDE. Para obter estes serviços, o módulo tem uma *thread* que periodicamente executa a operação *GetRecords* do catálogo, solicitando os registros referentes a serviços que foram incluídos ou alterados após a última leitura. Para cada serviço recuperado na consulta, o módulo executa a operação *GetRecordById*, para recuperar o seu registro de metadados;
- **aquisição de *feature types*:** este módulo é responsável por obter as informações sobre todos os *feature types* oferecidos por cada serviço identificado pelo módulo de aquisição de serviços. Para isto, o módulo executa a operação *GetCapabilities* de cada serviço recuperado pelo módulo de aquisição de serviços e obtém o seu respectivo documento de funcionalidades;
- **anotação:** depois que as informações sobre os novos serviços e seus respectivos *feature types* são recuperadas, a próxima etapa do processo

de coleta de informações consiste em identificar as características espaciais, temáticas e temporais de cada *feature type*. Esta tarefa é realizada pelo módulo de anotação, que é subdividido em três módulos: anotação espacial, temática e temporal; e

- **persistência:** este módulo é responsável pelo armazenamento persistente de todas as informações obtidas, a partir das etapas anteriores, na base de dados usada pelo SESDI.

A seguir são mostrados detalhes sobre o módulo de anotação, que é o principal módulo do subsistema de coleta de informações.

5.2 A anotação espacial

O primeiro tipo de anotação realizada durante o processo de coleta de informações é a anotação espacial, que tem como objetivo identificar a região geográfica coberta por cada *feature type* oferecido por um serviço.

Atualmente, informações sobre a extensão geográfica coberta por serviços de dados geográficos podem ser obtidas a partir de duas fontes: o seu registro de metadados e o seu documento de funcionalidades. No registro de metadados, um elemento é usado para prover este tipo de informação, que é normalmente descrita através das coordenadas do *bounding-box* da região coberta pelo serviço. Na Figura 5.1, é mostrado um exemplo da definição da extensão geográfica de um serviço no padrão ISO 19115. A figura mostra, na respectiva ordem, as coordenadas do menor ponto ao oeste, o maior ponto ao leste, o menor ponto ao sul e o maior ponto ao norte.

Outra fonte de informações sobre a extensão geográfica do serviço é o seu documento de funcionalidades. Neste documento, a extensão geográfica também é descrita através das coordenadas do seu *bounding-box*. Entretanto, diferentemente do registro de metadados, no qual uma única extensão geográfica é utilizada para descrever todo o serviço, este documento possui elementos que descrevem a extensão geográfica de cada *feature type* oferecido pelo serviço. Na Figura 5.2, é mostrada a descrição de uma camada no documento de funcionalidades de um serviço. A sua extensão geográfica é descrita pelo componente *EX_GeographicBoundingBox*.

```
- <extent>
  - <EX_Extent>
    - <geographicElement>
      - <EX_GeographicBoundingBox>
        - <westBoundLongitude>
            <gco:Decimal>-17.3</gco:Decimal>
          </westBoundLongitude>
        - <eastBoundLongitude>
            <gco:Decimal>51.1</gco:Decimal>
          </eastBoundLongitude>
        - <southBoundLatitude>
            <gco:Decimal>-34.6</gco:Decimal>
          </southBoundLatitude>
        - <northBoundLatitude>
            <gco:Decimal>38.2</gco:Decimal>
          </northBoundLatitude>
        </EX_GeographicBoundingBox>
      </geographicElement>
    </EX_Extent>
  </extent>
```

Figura 5.1: Exemplo de extensão geográfica no serviço de catálogo

Como as informações providas pelo documento de funcionalidades do serviço descrevem as regiões geográficas de cada *feature type* de uma forma mais detalhada quando comparadas com as informações oferecidas pelo registro de metadados, optou-se pela utilização das extensões geográficas oferecidas por este documento como base para a anotação espacial das camadas do SESDI. Tal opção foi importante para evitar os problemas de cobertura e precisão discutidos no Capítulo 1. Assim, durante o processo de anotação, o módulo de anotação espacial processa as informações do documento de funcionalidades do serviço, e extrai o *bounding-box* que representa a região geográfica referente a cada camada oferecida pelo serviço.

```
<Layer queryable="1">
  <Name>1</Name>
  <Title>Water Monitoring Locations</Title>
  <Abstract>Water Monitoring Locations</Abstract>
  <CRS>CRS:84</CRS>
  <CRS>EPSG:4326</CRS>
  <CRS>EPSG:4269</CRS>
- <EX_GeographicBoundingBox>
    <westBoundLongitude>-160.143988</westBoundLongitude>
    <eastBoundLongitude>-64.579847</eastBoundLongitude>
    <southBoundLatitude>17.679675</southBoundLatitude>
    <northBoundLatitude>49.321898</northBoundLatitude>
  </EX_GeographicBoundingBox>
  <BoundingBox CRS="CRS:84" minx="-160.143988" miny="17.679675" maxx="-64.579847" maxy="49.321898"/>
  <BoundingBox CRS="EPSG:4326" minx="17.679675" miny="-160.143988" maxx="49.321898" maxy="-64.579847"/>
  <BoundingBox CRS="EPSG:4269" minx="17.679675" miny="-160.143988" maxx="49.321898" maxy="-64.579847"/>
+ <Style></Style>
</Layer>
```

Figura 5.2: Exemplo de extensão geográfica no documento de funcionalidades

5.3 A anotação temática

O processo de anotação temática consiste em tentar identificar o significado dos dados oferecidos por cada *feature type* provido por um serviço. Para permitir que este significado possa ser compreendido e explorado pelo motor de busca do SESDI, o módulo de anotação temática tenta associar cada camada do serviço a um ou mais conceitos definidos nas ontologias cadastradas na base de dados do arcabouço.

Atualmente, informações sobre o tema do serviço podem ser encontradas tanto no seu registro de metadados quanto no seu documento de funcionalidades. No seu registro de metadados, estas informações são descritas através de um conjunto de palavras-chave, que contêm uma série de termos relacionados ao seu conteúdo. Um exemplo da descrição temática de um serviço de acordo com o padrão ISO 19115 é mostrado na Figura 5.3. No registro desta figura, pode-se notar que o serviço é descrito pelas palavras-chave *watersheds*, *river basins*, *water resources*, *hydrology*, *AQUASTAT* e *AWRD*.

```
<gmd:descriptiveKeywords>
 - <gmd:MD_Keywords>
   - <gmd:keyword>
       <gco:CharacterString>watersheds</gco:CharacterString>
     </gmd:keyword>
   - <gmd:keyword>
       <gco:CharacterString>river basins</gco:CharacterString>
     </gmd:keyword>
   - <gmd:keyword>
       <gco:CharacterString>water resources</gco:CharacterString>
     </gmd:keyword>
   - <gmd:keyword>
       <gco:CharacterString>hydrology</gco:CharacterString>
     </gmd:keyword>
   - <gmd:keyword>
       <gco:CharacterString>AQUASTAT</gco:CharacterString>
     </gmd:keyword>
   - <gmd:keyword>
       <gco:CharacterString>AWRD</gco:CharacterString>
     </gmd:keyword>
   + <gmd:type></gmd:type>
  </gmd:MD_Keywords>
</gmd:descriptiveKeywords>
```

Figura 5.3: Descrição temática em um registro de metadados

No documento de funcionalidades, a descrição temática do serviço também é realizada através de um conjunto de palavras-chave. Entretanto, assim como na descrição espacial, o documento fornece palavras-chave específicas para cada *feature type*. Desta forma, foi definido que a identificação dos conceitos usados para a anotação

temática de cada camada seria realizada com base nas palavras-chave contidas neste documento. Esta escolha deu-se pelo fato de que este documento fornece palavras-chave específicas para cada *feature type* oferecido pelo serviço. A desvantagem desta abordagem é que, em muitos casos, estas palavras-chave são omitidas no documento. Entretanto, quando isto acontece, a anotação pode ser realizada a partir de outros atributos, como o seu título ou a descrição textual.

Uma abordagem baseada na *Wikipedia*

Para resolver o problema da anotação temática de *feature types*, foi desenvolvida uma abordagem que usa a *Wikipedia*[1] como ferramenta de apoio. Esta escolha ocorreu pelo fato de que a *Wikipedia* oferece uma série de vantagens para a realização deste tipo de tarefa, tais como:

- **Resolução de sinônimos:** a *Wikipedia* consegue resolver diversas consultas envolvendo sinônimos. Nesta enciclopedia, quando o usuário entra com uma consulta por um verbete que representa um sinônimo do título de alguma página existente em sua base de dados, a consulta é automaticamente redirecionada para esta página. Por exemplo, uma consulta pelo verbete *Waterbody* é automaticamente redirecionada para a página *Body Of Water*;
- **Resolução de acrônimos:** a *Wikipedia* consegue resolver muitas requisições envolvendo acrônimos, que são recursos usados frequentemente para a descrição de dados geográficos. Por exemplo, caso o usuário realize uma busca pelo verbete *EPA*, o sistema automaticamente a redireciona para a página *United States Environmental Protection Agency,* que corresponde à agência federal de meio ambiente norte-americana. Caso o acrônimo requisitado pela consulta tenha mais de uma utilização, o sistema mostra uma página de desambiguação contendo todas as suas utilizações;
- **Categorização:** inicialmente, a anotação temática de *feature types* pode ser vista como um problema de classificação, que é comumente abordado por técnicas de aprendizado de máquina, tais como redes neurais e máquinas de vetores de suporte. Entretanto, algumas características deste

[1] http://www.wikipedia.org

domínio dificultam a sua solução através destas técnicas tradicionais. Algumas destas características incluem o grande número de classes usadas para a classificação (que pode chegar a milhares), que faz com que muitos modelos tenham que ser gerados e verificados durante a classificação, e a disponibilidade de textos muito curtos (algumas vezes poucas palavras) para ser usado como base para a classificação, o que dificulta a geração de bons modelos e bons conjuntos de treinamento. A vantagem oferecida pela *Wikipedia* é que todas as suas páginas já são classificadas em uma ou mais categorias. Por exemplo, a página *Shoreline* mostra que este tema pertence às categorias *Landform*, *Coastal Geography* e *Topography Stubs*. Destarte, depois que a página referente ao *feature type* é encontrada, o problema de classificação se reduz ao problema de associar as categorias da página aos conceitos usados pelas ontologias, que é bem menos complexo do que o problema de classificação;

- **Identificação de palavras flexionadas e relacionadas:** outra vantagem oferecida pela *Wikipedia* é a capacidade de resolver consultas mesmo que haja diferenças entre o título requisitado e o título da página devido a flexões de palavras, como plural ou verbos conjugados. Por exemplo, uma consulta pelo título *Flooding* é automaticamente redirecionada para a página *Flood.* Outra característica importante desta ferramenta é a sua capacidade de resolver algumas consultas por palavras relacionadas. Por exemplo, uma busca pelo verbete *Carbon Flux* é automaticamente redirecionada para a página *Carbon dioxide in Earth's atmosphere*.

A partir deste ponto, nesta subseção, o termo *título* será usado para se referir ao texto usado como referência para a anotação temática. Esta escolha se deu porque os exemplos que serão abordados para ilustrar este processo correspondem a anotações geradas a partir deste tipo de informação. Entretanto, é importante ter em mente que o processo de anotação temática de uma camada pode ser realizado com base nas suas palavras-chave, no seu título ou em sua descrição textual.

A abordagem usada para a anotação temática é composta por três etapas: seleção das páginas, identificação dos conceitos e a seleção dos conceitos usados para a anotação. Estas etapas são explicadas em mais detalhes nas próximas subseções.

Selecionando a página

A primeira etapa delas consiste em localizar, na *Wikipedia*, uma página que represente o conteúdo da informação oferecida pelo *feature type* que está sendo anotado. Para isto, o módulo de anotação temática envia uma consulta para a enciclopédia, passando como parâmetro o título do *feature type*. Caso a pesquisa retorne algum resultado, a página retornada é selecionada para a anotação e o processo segue para a próxima etapa. Por exemplo, durante a anotação da camada "*Congaree National Park*", uma consulta pelo seu título já retorna uma página, com o mesmo nome, correspondente ao conteúdo oferecido pela enciclopédia.

Entretanto, existem casos nos quais a busca direta pelo título do *feature type* não recupera nenhuma página. Quando isto acontece, o módulo realiza novas consultas com base na combinação dos *tokens* que compõem o título. Para isto, o título é processado por uma ferramenta de análise de partes do discurso, chamada *TreeTagger*[2], e as suas *stop words,* que são elementos que não possuem grande influência para a resolução das consultas, tais como artigos, numerais, pronomes e preposições, são removidas. Após realizar este processamento, o módulo gera novas consultas a partir das combinações com os *tokens* restantes.

Um busca por cada combinação gerada é enviada para a *Wikipedia* e as páginas obtidas a partir destas combinações são usadas para a anotação. Por exemplo, durante a anotação do *feature type* "*National Park Points*", a busca pelo título não retorna nenhuma página. Entretanto, quando uma busca é feita pelo título "*National Park*", formado a partir da combinação dos dois primeiros *tokens* do título original, a página referente a esta camada é localizada. Caso não se consiga obter uma página para nenhuma das combinações, o módulo envia uma requisição por cada *token* que compõe o título do *feature type*. Por exemplo, uma busca pelo título "*Park Boundary*" não retorna nenhuma página. Entretanto, quando os temas "*Park*" e "*Boundary*" são pesquisados separadamente, são encontradas páginas para ambos os temas.

Em algumas situações, mais de uma página pode ser recuperada. Este tipo de situação pode acontecer quando o título requisitado possui mais de um significado. Por exemplo, uma busca pelo título "*Europe*", mostra que o mesmo pode se referir, dentre vários outros significados, ao continente europeu, a uma banda de rock sueca e a um jornal. Quando este tipo de situação ocorre, meios são necessários para identificar,

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

dentre todas as páginas recuperadas, aquelas que se referem ao *feature type* que está sendo anotado. Nestes casos, para realizar a seleção das páginas, o módulo analisa a presença dos *tokens* que compõem o título dentro de cada página selecionada. Caso o número de *tokens* presentes no título seja inferior a cinco, as palavras-chave usadas para a descrição do serviço que oferece a camada que está sendo anotada são usadas para complementar o texto da consulta.

A busca é realizada através do modelo vetorial, que é um modelo clássico usado para a recuperação de documentos. A consulta é representada por um vetor n-dimensional, no qual cada dimensão representa um dos *tokens* usado para a consulta. Para cada dimensão do vetor da consulta é atribuído o valor 1. Da mesma forma que a consulta, cada página também é representada como um vetor. Neste caso, para cada uma de suas dimensões são atribuídos os valores 1, quando a página contém o *token* referente a esta dimensão, ou 0, quando a página não contém tal *token*. Estes pesos binários foram escolhidos porque, pela observação dos experimentos iniciais, pode-se perceber que as páginas irrelevantes não possuem a ocorrência da maior parte dos *tokens* usados na consulta. Este fato, aliado ao grande número de *tokens* usados durante a consulta, inviabiliza a utilização das métricas *tf* e *idf*, para computar o peso destas dimensões, como normalmente ocorre neste tipo de consulta.

Depois que os vetores que representam a página e a consulta são gerados, o cosseno do ângulo formado por estes vetores é usado para determinar a relevância de cada página. Depois que todas as páginas são processadas, aquelas que apresentam o maior grau de relevância são selecionadas para a realização da próxima etapa do processo de anotação.

A Tabela 5.1 mostra um exemplo do processo de seleção de páginas durante a anotação temática de um *feature type* chamado *Watersheds*. Neste caso, a consulta é complementada com alguns *tokens* (*river, basin, hydrology, water, resource*) providos pelo registro de metadados. Esta tabela mostra as dimensões dos vetores usados para representar a consulta e algumas das páginas retornadas pela *Wikipedia* após uma busca pelo título desta camada. A análise da tabela mostra que, ao final desta consulta, o módulo de anotação opta pela página *Drainage Basin*, que apresenta a maior relevância.

A Tabela 5.2 mostra alguns exemplos de páginas que foram selecionadas pelo módulo de anotação durante a primeira etapa do processo de anotação temática de algumas camadas oferecidas por serviços de dados geográficos que compõem o estudo de caso usado pelo SESDI. Nesta tabela, a primeira coluna representa o nome da

camada, enquanto que a segunda representa o título da página que foi selecionada pelo módulo para a realização da sua anotação.

Tabela 5.1: Exemplo de seleção de páginas

Página	Dimensões do Vetor						Relevância
	watershed	river	basin	hydrology	water	resource	
Drainage basin	1	1	1	1	1	0	0,9128
Watershed Stroke	1	0	0	0	1	1	0,7071
Watershed (Band)	1	0	0	0	1	0	0,5773
Watershed (television)	1	0	0	0	0	0	0,4082
Watershed (Opeth Album)	1	0	0	0	0	0	0,4082

Tabela 5.2: Exemplos de páginas selecionadas

Nome do Feature Type	Página Selecionada na Wikipedia
Rivers	River
Watershed	Drainage Basin
Congaree National Park	Congaree National Park
Toxic Releases	Toxicity
Superfund Sites	List of Superfund sites in the United States

Identificando os conceitos

Depois de selecionar as páginas que melhor representam o conteúdo do *feature type* que está sendo anotado, a próxima etapa do processo de anotação temática consiste em identificar, dentre as ontologias disponíveis, conceitos que possam ser usados para a anotação do *feature type*. Para isto, o módulo inicialmente identifica os conceitos que podem ser relacionados para cada categoria a qual a página está associada. Esta identificação é feita através do casamento entre os nomes da categoria e o nome dos

conceitos das ontologias, e todos os conceitos cujo nome é igual ao da categoria que está sendo processada são pré-selecionados.

A desvantagem do método de casamento de *strings* é que podem ser selecionados conceitos cujos nomes casam com o nome da categoria, mas que não são relacionados à mesma. Este tipo de situação acontece quando um conceito é ambíguo e pode ser usado em mais de um domínio de aplicação. Isto requer um mecanismo que permita avaliar se cada conceito selecionado se refere ou não à categoria que está sendo processada.

Para realizar a validação de um conceito, o módulo recupera, junto à *Wikipedia,* a página da categoria que está sendo processada. A importância desta página para a validação dos conceitos é que a ela possui informações importantes sobre a semântica da categoria. Esta semântica é descrita através de três tipos de informação: as categorias as quais a categoria está relacionada, as suas subcategorias e os nomes de todas as páginas que estão associadas à mesma. Por exemplo, a página referente à categoria *Hydrology*, que é uma das categorias da página *Drainage Basin*, mostra que esta categoria está relacionada às categorias *Physical geography*, *Water*, e *Geophysics,* e possui como subcategorias: *Aquifers*, *Hydrogeology* e *Hydrology Models*, e que páginas como *Hydrography*, *Flood* e *Drainage Model* são associadas a esta categoria.

Depois de localizar a página da categoria, o módulo executa uma consulta para verificar se os nomes dos conceitos que compõem o domínio de aplicação do conceito que está sendo validado, podem ser encontrados nesta página. Esta consulta também é realizada através do modelo vetorial, usando exatamente a mesma abordagem na seleção de páginas. Durante esta etapa, apenas os conceitos cuja relevância no seu domínio de aplicação é superior a um *threshold* (limiar) pré-definido são aceitos e usados na anotação do *feature type*. Durante a implementação do módulo de anotação semântica, o valor 0.35 foi definido para este *threshold*. Este valor foi determinado através da realização de vários experimentos, que observavam a qualidade das anotações geradas para cada camada. Estes experimentos mostraram que, quando o valor deste limite é muito baixo, o módulo acaba gerando muitas anotações indevidas, o que prejudica a precisão das consultas temáticas. Da mesma forma, pode-se perceber que quando o inverso ocorre, muitas anotações relevantes acabam sendo rejeitadas, o que prejudica a cobertura destas consultas.

O resultado desta etapa é um conjunto de conceitos pré-selecionados após o processamento das categorias de todas as páginas selecionadas na primeira etapa. Por

exemplo, após o processamento da página "Drainage Basin", são pré-selecionados os conceitos "*Drainage Basin*", "*Landform*", "*Fluvial Landform*", "*Hydrology*" e "*River*".

Selecionando os conceitos

Finalmente, a última etapa do processo de anotação temática consiste em selecionar, dentre todos os conceitos pré-selecionados na etapa anterior, aqueles que serão usados para a descrição do *feature type*. Para isto, inicialmente, o módulo verifica se o conjunto de conceitos possui conceitos semanticamente relacionados entre si. Este tipo de situação ocorre caso o conjunto tenha um conceito que seja subsumido por outro conceito no mesmo conjunto.

Um exemplo deste tipo de situação ocorre caso o conjunto tenha, em sua composição, um conceito que seja uma subclasse do outro conceito. Quando este tipo de situação ocorre, o módulo mantém apenas o conceito mais específico, descartando o conceito que é mais geral. Esta escolha é feita porque o conceito mais específico oferece uma informação mais detalhada a respeito do *feature type*. Um exemplo desta situação acontece durante o processamento do conjunto mostrado na etapa anterior. Ao analisar os conceitos pré-selecionados, o módulo de anotação temática identifica que o conceito "*FluvialLandform*" corresponde a uma especialização do conceito "*Landform*". Destarte, o conceito "*Landform*" é removido do conjunto e descartado do resultado final.

O resultado do processo de anotação temática é um conjunto de conceitos que descrevem o tema referente à informação oferecida pelo *feature type*. Estas anotações são armazenadas na base de dados do SESDI e usadas para a resolução de consultas com restrições de tema.

5.4 A anotação temporal

A tarefa da anotação temporal consiste em determinar o intervalo temporal referente a um *feature type*. As informações sobre a extensão temporal dos dados oferecidos por um serviço podem ser encontradas no registro de metadados. A forma como esta informação é descrita depende do padrão adotado pela IDE. No padrão ISO 19115 (ISO, 2003), que é o padrão desenvolvido pela ISO para a especificação de metadados geográficos e que serve como base para vários outros padrões de metadados, a referência temporal de um recurso é geralmente descrita através de um intervalo temporal. Tal intervalo é definido através de dois atributos chamados *beginPosition* e

endPosition, que definem, respectivamente, os seus limites inicial e final. O valor destes atributos é comumente descrito no formato do padrão ISO 8601 (ISO, 2004). Na Figura 5.4, é mostrada a descrição da extensão temporal de um serviço no padrão ISO 19115. Os valores dos atributos *beginPosition* e *endPosition* mostram que os dados oferecidos por este serviço se referem ao período de 01 de janeiro de 2000 a 08 de janeiro de 2008.

```
<gmd:extent>
    <gmd:EX_Extent>
        <gmd:temporalElement>
            <gmd:EX_TemporalExtent>
                <gmd:extent>
                    <gml:TimePeriod gml:id="timeperiod1">
                        <gml:beginPosition>2000-01-01T04:29:00</gml:beginPosition>
                        <gml:endPosition>2008-01-08T04:29:00</gml:endPosition>
                    </gml:TimePeriod>
                </gmd:extent>
            </gmd:EX_TemporalExtent>
        </gmd:temporalElement>
    </gmd:EX_Extent>
</gmd:extent>
```

Figura 5.4: Extensão temporal de um recurso no serviço de catálogo

As limitações para identificar a extensão temporal de um recurso são ainda maiores no documento de funcionalidades do serviço, que não oferece qualquer atributo específico para descrever este tipo de informação. Devido a estas limitações, o processo de anotação temporal de *feature types* foi dividido em duas etapas. Na primeira, chamada de anotação do serviço, o módulo de anotação temporal processa as informações dos metadados para tentar identificar o intervalo temporal correspondente ao serviço como um todo. Na segunda etapa, chamada de anotação de *feature type*, o componente processa as informações contidas no documento de funcionalidades para tentar identificar um intervalo temporal mais específico para cada *feature type*. A seguir, descrevemos cada uma destas etapas.

5.4.1 A anotação temporal em nível de serviços

A anotação temporal do serviço corresponde ao processo de identificar o intervalo temporal referente a todo o seu conjunto de dados. Para identificar esta informação, o anotador temporal analisa as informações contidas no registro de metadados do serviço. Durante este processo, o intervalo de referência do serviço é identificado através de dois atributos existentes neste registro: a extensão temporal e a

frequência de atualização do serviço. A extensão temporal é usada para identificar o período temporal referente ao conjunto de dados oferecido pelo serviço. Por outro lado, a frequência de atualização é usada para identificar se o serviço continua sendo continuamente atualizado. Caso isto aconteça, o modelo considera que o serviço é persistente e, consequentemente, não possui um limite final. No modelo usado pelo SESDI, são considerados persistentes os serviços que possuem os seguintes valores para a frequência de atualização: *annually, continually, daily, monthly* e *weekly*.

Quando a extensão temporal do serviço não está presente nos seus metadados, o *crawler* precisa então identificar a extensão temporal a partir dos valores de outros atributos. Tal tarefa é executada de acordo com uma lista de atributos, organizados por ordem de prioridade. Caso o atributo consultado contenha esta informação, o seu valor é extraído, normalizado para um intervalo temporal e associado ao serviço, encerrando o processo de anotação. Caso contrário, o próximo atributo é consultado e o processo é repetido até que as informações temporais sejam encontradas ou até que todos os atributos tenham sido consultados. Os atributos consultados para obter a informação temporal de um serviço, em ordem de prioridade, são as suas palavras-chave, o seu título e a sua descrição textual. Finalmente, caso as informações temporais não possam ser encontradas em nenhum destes atributos, a data de inclusão do registro de metadados é usada como a referência temporal do serviço.

Dentre os atributos consultados durante o processo de anotação, alguns são representados diretamente na forma de datas, como a extensão temporal e a data da inclusão do registro. Quando a informação temporal é obtida a partir destes atributos, a sua extração é bastante simples. Para isto, é necessário apenas converter o valor do atributo para um intervalo temporal. Caso o valor do atributo seja um intervalo, este é convertido para o formato da data usado para a representação dos dados, através de um processo de normalização.

O processo de normalização de datas varia de acordo com o formato das informações temporais. Caso o valor do atributo seja apenas um ponto no tempo, como por exemplo, 2010, a sua conversão para um intervalo pode ser feita de duas formas, dependendo da sua frequência de atualização. Caso o serviço seja continuamente atualizado, o valor é convertido para um intervalo persistente, que possui como limite final o valor *now*, que significa que o limite final corresponde ao instante de tempo em que a consulta é realizada. Caso contrário, o valor é normalizado para um intervalo contendo os limites do valor encontrado para o atributo. A granularidade deste intervalo

A normalização das anotações temporais de um texto depende do número de anotações encontradas. Caso o arquivo tenha apenas uma anotação temporal, o seu valor é normalizado da mesma forma que os três últimos exemplos da Tabela 5.3. Caso o arquivo tenha mais de uma anotação temporal, o intervalo do *feature type* corresponde ao menor intervalo que engloba todas as anotações temporais encontradas. Por fim, a ausência de anotações temporais significa que não foi encontrado nenhum elemento temporal no texto que foi analisado. Tal situação indica que a referência temporal não pode ser obtida através deste atributo.

5.4.2 A anotação temporal em nível de *feature types*

Visando localizar informações mais precisas com relação à extensão temporal, o modelo usado por nosso motor de busca tenta identificar informações temporais específicas para cada *feature type* oferecido por um serviço. Entretanto, diferentemente dos serviços, a anotação temporal de um *feature type* é realizada com base nas informações contidas no documento de funcionalidades do serviço. Tal documento é obtido invocando-se a operação *GetCapabilities* do serviço que está sendo processado.

Uma característica importante do documento de funcionalidades é que, diferentemente do registro de metadados, o mesmo não tem quualquer atributo especifico para definir as informações temporais dos *feature types* oferecidos pelo serviço. Assim, a anotação destes elementos precisa ser feita através da extração de expressões temporais presentes nos valores de alguns de seus atributos. Para realizar esta tarefa, são consultados, por ordem de prioridade e para cada *feature type*, as suas palavras-chave, o seu título e a sua descrição textual.

Como todos os atributos usados para a anotação de *feature types* são textuais, as informações temporais presentes nestes elementos precisam ser extraídas através do processamento dos textos referentes aos seus valores. O procedimento adotado aqui para a realização desta tarefa é o mesmo utilizado durante a anotação de serviços. Os valores obtidos após este processo são usados como a extensão temporal da camada que está sendo processada. Caso a extensão temporal do *feature type* não possa ser obtida a partir de nenhum dos atributos verificados, assume-se que a sua extensão temporal é a mesma obtida para o seu respectivo serviço.

5.5 Considerações finais

Este capítulo apresentou a especificação do SESDI, o arcabouço proposto por esta tese para melhorar a recuperação de informações geográficas em IDEs. Inicialmente, o capítulo abordou a especificação do arcabouço, discutindo os requisitos levantados, o modelo proposto para a representação e armazenamento dos dados e o projeto arquitetural. Depois, o capítulo discutiu como as informações usadas pelo arcabouço são obtidas a partir da IDE. Finalmente, foi discutido como as informações espaciais, temáticas e temporais de cada *feature type* são identificadas por cada módulo de anotação do arcabouço. O próximo capítulo mostra como as informações obtidas através da coleta e anotação de serviços e *feature types* são usadas durante o processo de recuperação da informação.

Capítulo 6 – SESDI – Recuperação baseada em ranking

Este capítulo mostra como as informações obtidas através do processo de coleta de informações são usadas para o processo de resolução de consultas. O capítulo é dividido em quatro partes. Na seção 6.1, é discutida uma métrica para avaliar o *ranking* espacial, enquanto que na seção 6.2 a abordagem proposta para avaliar o *ranking* temático é abordada. A seção 6.3 discorre sobre o *ranking* temporal. Finalmente, a seção 6.4 apresenta uma métrica que combina todas as métricas anteriores para avaliar o *ranking* global.

6.1 *Ranking* espacial

Atualmente, muitas ferramentas de busca espacial recuperam dados geográficos com base no seu *bounding-box*. Geralmente, estas ferramentas recuperam, durante o processo de recuperação de informações, todos os dados cuja geometria intersecta a região geográfica definida na consulta do usuário. O problema deste tipo de solução é que a mesma considera que todos os dados recuperados durante uma consulta têm a mesma relevância para o usuário. Tal característica é indesejável, principalmente porque este tipo de consulta normalmente retorna uma grande quantidade de recursos. Por exemplo, a realização de uma consulta na base de dados usada como estudo de caso do arcabouço, por camadas cuja região geográfica intersecta a região geográfica do estado de *New Jersey,* retorna um percentual de aproximadamente 84% das camadas cadastradas. Nestes casos, a ordem em que os resultados são mostrados para o usuário é normalmente estabelecida através da ordem em que os mesmos são recuperados ou através da ordenação de acordo com algum dos atributos de cada resultado, como nome ou título. De toda forma, neste tipo de consulta, um objeto cuja região geográfica é idêntica à região da consulta, acaba tendo a mesma relevância que outro objeto cuja região geográfica intersecta apenas uma pequena parte da mesma.

A limitação imposta por este tipo de consulta faz surgir a necessidade de uma medida que permita avaliar o grau de similaridade existente entre os *bounding-boxes* de dois objetos. O objetivo desta medida é avaliar o quanto um determinado *feature type* é

relevante para a consulta do usuário, considerando apenas a dimensão espacial de ambos.

A primeira etapa da resolução de consultas espaciais consiste em selecionar os *feature types* que satisfazem as restrições definidas na requisição do usuário. Para a resolução deste tipo de consulta, o SESDI oferece suporte para a recuperação de informação baseada em três tipos de relacionamentos topológicos: *intersects*, que recupera todos os *feature types* cuja região geográfica intersecta a região da consulta; *covers*, que recupera todos os *feature types* cuja região geográfica cobre totalmente esta região; e *within*, que recupera todas as camadas cuja região geográfica está totalmente contida dentro da região da consulta. Depois de selecionar as camadas que são relevantes, a próxima etapa consiste em calcular o *ranking* para cada *feature type* recuperado. Este cálculo é baseado em duas métricas: o grau de sobreposição das duas regiões e o grau de relevância espacial. Porém, antes de mostrar como estas duas métricas são obtidas, esta seção discute os requisitos que nortearam o desenvolvimento das mesmas.

6.1.1 Os requisitos para o *ranking* espacial

Antes de definir a medida de *ranking* que seria usada para a resolução de consultas espaciais, foi definido que a mesma deveria satisfazer os seguintes requisitos:

- **Assimetria:** dadas duas regiões geográficas R e S, o valor do *ranking* espacial para estas regiões deve ser dependente da ordem em que as mesmas estivessem sendo comparadas. Para compreender este requisito, vamos supor uma consulta na qual o usuário procura por mapas sobre o estado da Paraíba. Caso um *feature type* cubra toda a região Nordeste do Brasil, a consulta do usuário é totalmente resolvida, uma vez que a região geográfica requisitada na consulta é completamente coberta pela região do *feature type*. Agora, vamos supor o caso inverso, ou seja, quando o usuário procura por mapas do Nordeste e o *feature type* cobre apenas o estado da Paraíba. Neste caso, a consulta do usuário não é totalmente resolvida, uma vez que o *feature type* cobre apenas uma parte da região solicitada. Para representar este tipo de situação, na primeira consulta, o valor do *ranking* espacial entre as duas regiões deve ser considerado maior do que no segundo caso;

- **A similaridade das regiões deve ser considerada:** a avaliação do *ranking* entre duas regiões geográficas deve levar em consideração o grau de semelhança entre as duas regiões. O objetivo deste requisito é priorizar *feature types* associados a regiões geográficas mais parecidas com a região definida na consulta. Este requisito satisfaz a Lei de Tobler para a geografia, que define que "*no mundo, todas as coisas se parecem; mas coisas mais próximas são mais parecidas que aquelas mais distantes*" (TOBLER, 1970).
- **A relevância geográfica deve ser considerada:** outra informação importante para avaliar a relevância de um *feature type* para uma consulta seria o grau de importância que a região geográfica definida na consulta tem para o serviço que o oferece. A importância de uma região geográfica para um determinado serviço deve ser avaliada através da análise da região geográfica de cada *feature type* oferecido pelo mesmo. Com esta medida, os dados oferecidos por serviços que disponibilizam mais informações sobre a região consultada deveriam obter um valor de *ranking* maior do que aqueles oferecidos por serviços para os quais esta região é menos relevante.

6.1.2 O grau de sobreposição espacial

O grau de sobreposição foi proposto para avaliar a similaridade entre a região geográfica definida na consulta e a região coberta por um *feature type* sob avaliação. Visando satisfazer os requisitos definidos para o *ranking* espacial, para o cálculo desta métrica, foi usada a ideia desenvolvida por Tversky (1977) para avaliar a similaridade entre dois objetos. Esta medida de similaridade tem como característica principal o fato de que a similaridade entre os dois objetos que estão sendo comparados é calculada levando-se em consideração tanto as características que eles possuem em comum quanto aquelas em que os mesmos são diferentes. A Equação 1 mostra a adaptação da equação de Tversky para avaliar a similaridade espacial. Nesta equação, A é a região geográfica requisitada pela consulta do usuário e B é a região coberta pelo *feature type* que está sendo avaliado.

$$overlap(A,B) = \frac{area(A \cap B)}{area(A \cap B) + \alpha * area(A - B) + (1-\alpha) * area(B - A)} \quad (1)$$

Onde:

- *area (A ∩ B)* representa o valor da área que as regiões geográficas compartilham entre si;
- *area (A - B)* representa o valor da área que pertence à região A, mas que não pertence à região B;
- *area (B – A)* representa o valor da área que pertence à região B, mas que não pertence à região A;
- A constante α representa o peso que os complementos de cada região possuem durante o cálculo da similaridade.

Para satisfazer o requisito de assimetria, o valor 0.87 foi usado para a constante α. Este valor foi determinado através de uma técnica chamada *weighting* (FOX; SHAW, 1993), que permite avaliar o melhor peso para um determinado critério em um processo de tomada de decisão. Na primeira etapa deste procedimento, um conjunto de treinamento foi gerado, a partir de várias consultas espaciais. Este conjunto é uma quádrupla (a, b, c, d), no qual *a* representava o valor de *area (A ∩ B)*, *b* representava o valor de *area (B – A)*, *c* representava o valor de *area (B-A)* e *d* representava o valor esperado para o grau de sobreposição.

Depois que estes conjuntos de treinamentos eram gerados, foram calculados o peso que as variáveis *b* e *c* tinham no cálculo do grau de *overlap*. Esta relevância era determinada através de um método estatístico, chamado de coeficiente de correlação de Pearson, que é obtido através da Equação 2. Tal coeficiente é usado para estimar o quanto o valor de uma variável afeta o valor de outra variável. Na Equação 2, x representa a variável que está sendo observada, enquanto y representa o grau de sobreposição, enquanto os demais valores representam, respectivamente, os valores da variância destas duas variáveis.

$$Pearson(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}} \qquad (2)$$

Depois de encontrados os valores de correlação de *b* e *c* com relação ao grau de sobreposição, a última etapa para a obtenção dos pesos consistiu em normalizar os

valores destes coeficientes. Esta normalização foi realizada para ajustar os valores destes coeficientes para um valor entre 0 e 1. Terminada esta etapa, os valores normalizados para cada coeficiente foram usados como os pesos da equação. É importante destacar que as demais equações apresentadas durante o restante do capítulo envolvem a utilização de pesos. Os pesos de todas estas equações foram obtidos através da mesma técnica.

Alguns exemplos que ilustram os resultados obtidos para o grau de sobreposição são mostrados na Tabela 6.1. Esta tabela mostra as coordenadas geográficas dos *bounding-boxes* que cobrem diferentes localidades. A Tabela 6.2, por sua vez, mostra o grau de sobreposição existente entre estas regiões. Nesta tabela, as linhas representam as localidades requisitadas na consulta, enquanto as colunas representam as regiões cobertas pelo *feature type*.

Tabela 6.1: Exemplos de regiões geográficas

Região Geográfica	Bounding-Box
Amazônia	-73.9909, -18.041, -44.031, 5.2722
Brasil	-76.5126, -36.9484, -29.5852, 7.0460
Paraíba	-38.7651, -8.3029, -34.7934, -6.0269

Tabela 6.2: Tabela de similaridade entre bounding-boxes

	Amazônia	Brasil	Paraíba
Amazônia	1	0.7972	0
Brasil	0.3701	1	0.0050
Paraíba	0	0.0330	1

Através da análise dos resultados mostrados na Tabela 6.2, é possível perceber que o grau de sobreposição entre duas regiões idênticas sempre é igual a 1. Da mesma forma, pode-se perceber que a medida de similaridade é assimétrica. Por exemplo, caso o usuário esteja procurando por dados referentes à região amazônica e a região do *feature type* cobre todo o Brasil, o valor de similaridade é igual a 0.7972. Entretanto, quando o contrário acontece, o valor da similaridade diminui para 0.3701.

6.1.3 O grau de relevância espacial do serviço

Na recuperação de informação clássica, a frequência do termo t_i da consulta dentro de um documento d_j é comumente usada para determinar a relevância do documento em questão para a consulta. De uma forma semelhante, a avaliação do grau de importância que uma determinada região geográfica tem para um determinado serviço espacial pode melhorar a avaliação do *ranking* espacial. O objetivo desta medida é fazer com que *feature types* oferecidos por serviços cuja região geográfica definida em uma consulta espacial tenha um grau de relevância maior sejam mostrados primeiro no resultado final de uma consulta.

A avaliação da importância de um documento para a consulta é feita com base em uma série de métricas. Uma delas é chamada de *raw frequency* (*freq*), que corresponde ao número de vezes que o termo t_i aparece dentro de um documento d_j. A partir desta informação, pode-se calcular outra medida, chamada de frequência normalizada. A frequência normalizada (*tf*) corresponde à proporção entre a *raw frequency* e o número total de termos existentes no documento (n), como mostrado na Equação 3.

$$tf(i,j) = \frac{freq(i,j)}{n} \qquad (3)$$

Outra medida bastante popular neste tipo de recuperação da informação é chamada de frequência inversa (*idf*). O objetivo desta medida é avaliar o quanto um determinado termo é relevante para toda uma coleção de documentos. O seu valor é obtido através do logaritmo da proporção entre o número de documentos existentes no sistema (*N*) e o número de documentos do sistema nos quais o termo t_i aparece pelo menos uma vez (n_i), conforme é mostrado na equação 4.

$$idf(t_i) = \log\frac{N}{n_i} \qquad (4)$$

Por fim, o valor da importância de um determinado termo para um documento, pode ser avaliado através do produto da frequência normalizada pela frequência inversa, o que é definido pela equação 5. No modelo vetorial, este valor pode ser usado como o peso atribuído para a dimensão que representa o termo *ti*.

$$w_{i,j} = tf(i,j) * idf(i) \quad (5)$$

No SESDI, o grau de relevância espacial do serviço é a medida que indica o quanto a região geográfica definida na consulta é relevante para o serviço que oferece o *feature type* que está sendo avaliado. O seu valor é calculado através das mesmas métricas e equações usadas para avaliar a relevância de termos para documentos na recuperação da informação clássica. Contudo, algumas adaptações foram feitas para determinar os valores destas métricas a partir das informações armazenadas no modelo de dados usado por esta solução.

A primeira adaptação consiste em determinar o valor da frequência da região geográfica. Para computar o valor desta medida, foram considerados três tipos de abordagem. No primeiro tipo, a frequência de uma região geográfica *B* foi computada através do número de *feature types* cuja região geográfica era exatamente igual à região geográfica que estava sendo avaliada. Esta abordagem, apesar de ser facilmente implementada, apresentou algumas limitações. A principal delas era o fato de que os valores das coordenadas geográficas são especificados em números reais, usando diferentes precisões. Desta forma, o baixo número de regiões geográficas especificadas exatamente com as mesmas coordenadas fazia com que esta abordagem produzisse, na maior parte das vezes, valores muito baixos para a relevância espacial.

Em outro tipo de abordagem considerado, a frequência da região geográfica *B* foi computada através do número de *feature types* cuja região geográfica cobria completamente a região que estava sendo avaliada. Este tipo de abordagem amenizou o problema de precisão que ocorria no primeiro tipo de abordagem, uma vez que a frequência de *B* era computada sempre que a região geográfica do *feature type* cobria totalmente a região da consulta. Entretanto, o problema persistiu para os casos nos quais a região geográfica coberta pelo *feature type* fazia interseção à região da consulta, mas não a cobria totalmente.

O terceiro tipo de abordagem, que acabou sendo usado na métrica, consistiu em avaliar o grau de sobreposição entre a região definida na consulta e a região de cada *feature type* oferecido pelo serviço. Nesta abordagem, sendo *B* o *bounding-box* referente à região geográfica definida na consulta, a sua frequência é calculada através do somatório do grau de sobreposição entre a mesma e a região associada a cada *feature*

type oferecido pelo serviço. A escolha deste tipo de abordagem se deu porque a mesma oferece algumas vantagens com relação às demais. Estas vantagens acontecem porque, neste tipo, todas as interseções entre as regiões geográficas cobertas pelo serviço são consideradas. Assim, qualquer *bounding-box* relacionado à região definida pela consulta acaba influenciando no cálculo da frequência, independente do relacionamento topológico e da precisão dos valores de suas coordenadas. Ademais, o uso do grau de sobreposição permite que a frequência de uma região geográfica seja maior em serviços cujos *feature types* são associados a regiões geográficas mais parecidas com a mesma. Tal abordagem é descrita pela Equação 6. Nesta equação, Ti_{SPA} representa a região geográfica de um *feature type* Ti oferecido pelo serviço S.

$$freq(B,S) = \sum_{1}^{n} overlap(B, Ti_{SPA}) \quad (6)$$

Depois que a frequência é calculada, a próxima métrica a ser calculada é a frequência normalizada (*spa_tf*). Esta medida corresponde à proporção entre a frequência da região geográfica e o número de *feature types* oferecidos pelo serviço. Por fim, a frequência inversa (*spa_idf*) é calculada através do logaritmo da proporção entre o número de serviços cadastrados na base de dados da infraestrutura (*N*) e o número de serviços que oferecem pelo menos um *feature type* que cobre totalmente a região definida pela consulta (n_j).

Por fim, após avaliar estas medidas, é definido que, dados uma região geográfica B e um serviço S, o grau de relevância de B para S é dado pelo produto entre a frequência normalizada (*spa_tf*) e a frequência inversa (*spa_idf*). Este cálculo é mostrado através da equação 7.

$$relevance(B,S) = \frac{\sum_{i=1}^{n} overlap(B, Ti_{SPA})}{n} * \log \frac{N}{n_j} \quad (7)$$

Onde:

- *relevance* representa o grau de relevância da região geográfica definida pela consulta (B) para o serviço de dados geográfico (S);
- *overlap* representa o grau de sobreposição entre a região geográfica da consulta (B) e a região geográfica de um *feature type Ti* oferecido pelo serviço;

- n é o número de *feature types* oferecidos pelo serviço;
- N é o número de serviços de dados geográficos oferecidos pela infraestrutura;
- n_j é número de serviços da infraestrutura que possuem pelo menos um *feature type* cuja região geográfica cobre a região definida pela consulta.

6.1.4 Calculando o *ranking* espacial

Depois de descrever como os valores do grau de sobreposição e da relevância espacial seriam obtidos, foi necessário definir como os valores destas métricas seriam usados para determinar o *ranking* espacial de um *feature type* sob avaliação. Para determinar este valor, foram consideradas inicialmente duas métricas usadas pela recuperação da informação clássica: o cosseno angular e a distância euclidiana.

Na abordagem baseada no cosseno angular, a requisição do usuário era representada como um vetor bidimensional, de forma que a primeira dimensão representava o grau de sobreposição e a segunda representava o grau de relevância espacial. Para cada uma destas dimensões era atribuído o valor 1, que representava o valor máximo que poderia ser obtido para cada uma estas medidas. Os *feature types* também eram representados por vetores bidimensionais. Entretanto, nestes vetores, os valores das dimensões eram atribuídos a partir dos valores obtidos para cada uma destas métricas. Depois que estes vetores eram determinados, o valor da similaridade espacial de um *feature type* era calculado através do cosseno do ângulo formado entre o seu vetor e o vetor que representava a consulta. A desvantagem deste tipo de abordagem era que, independente dos valores das medidas usadas para a avaliação, os valores de similaridade obtidos eram muito próximos uns dos outros, o que dificultava a classificação dos resultados calculados.

Na abordagem baseada na distância euclidiana, a consulta do usuário era representada como um ponto em um espaço bidimensional. Neste espaço, os eixos representavam os valores do grau de sobreposição e o grau de relevância espacial. Assim como na abordagem baseada no cosseno angular, as consultas eram representadas pelo ponto $P = (1, 1)$, uma vez que 1 representa o valor máximo para cada métrica. Por outro lado, cada *feature type* era representado por um ponto (x, y), de forma que x e y representavam, respectivamente, os valores obtidos para o grau de sobreposição e para a relevância espacial. O *ranking* espacial era calculado através da distância euclidiana entre o seu ponto e o ponto que representava a consulta. A

desvantagem deste tipo de abordagem era que a similaridade era determinada através de qualquer número real. Entretanto, o valor de similaridade é mais intuitivo quando expresso através de um valor entre 0 e 1.

Devido às limitações das abordagens consideradas, optou-se pela utilização de uma outra abordagem, conhecida como soma ponderada (do inglês, *weighted sum*). Neste tipo de abordagem, que também é conhecida como análise de decisão com múltiplos objetivos, o valor de um objetivo é calculado como a soma de diversos critérios, no qual para cada critério é atribuído um peso. A vantagem deste método com relação aos demais é que o mesmo permite diferenciar melhor o valor da similaridade a partir dos critérios selecionados, permitindo uma melhor classificação dos resultados recuperados. Além disto, neste método, os pesos permitem que o valor da similaridade seja facilmente ajustado para um valor entre 0 e 1, tornando o valor da métrica mais intuitivo para o usuário.

De acordo com a métrica escolhida, dada uma região geográfica B definida na consulta e um *feature type* T oferecido pela infraestrutura, o *ranking* espacial de T pode ser calculado através da Equação 8. Nesta equação, T_{SPA} representa a região geográfica coberta por T, enquanto S representa o serviço pelo qual T é oferecido.

$$spa\_ranking(T,B) = w_1 * overlap(B,T_{SPA}) + w_2 * relevance(B,S) \quad (8)$$

Onde:

- *spa_ranking* representa a relevância do *feature type* T para a região geográfica requisitada pela consulta;
- *overlap* representa o grau de sobreposição entre a região requisitada e a região coberta pelo *feature type*;
- *relevance* representa a relevância que a região requisitada tem para o serviço pelo qual o *feature type* é oferecido.

A Figura 6.1 mostra os primeiros resultados obtidos através da resolução de uma consulta espacial no SESDI. No cenário desta consulta, o cliente deseja recuperar mapas sobre o estado de *New Jersey*. Este tipo de consulta retorna uma grande quantidade de resultados, uma vez que existem muitos *feature types* cuja extensão geográfica intersecta esta região. A análise desta figura mostra que o SESDI prioriza as camadas

cuja extensão geográfica é mais parecida com a região definida na consulta. Isto acontece devido ao grau de sobreposição. A figura mostra também que os primeiros resultados mostrados são oferecidos por serviços nos quais a região geográfica da consulta possui uma maior relevância. Tal opção ocorre devido ao grau de relevância espacial.

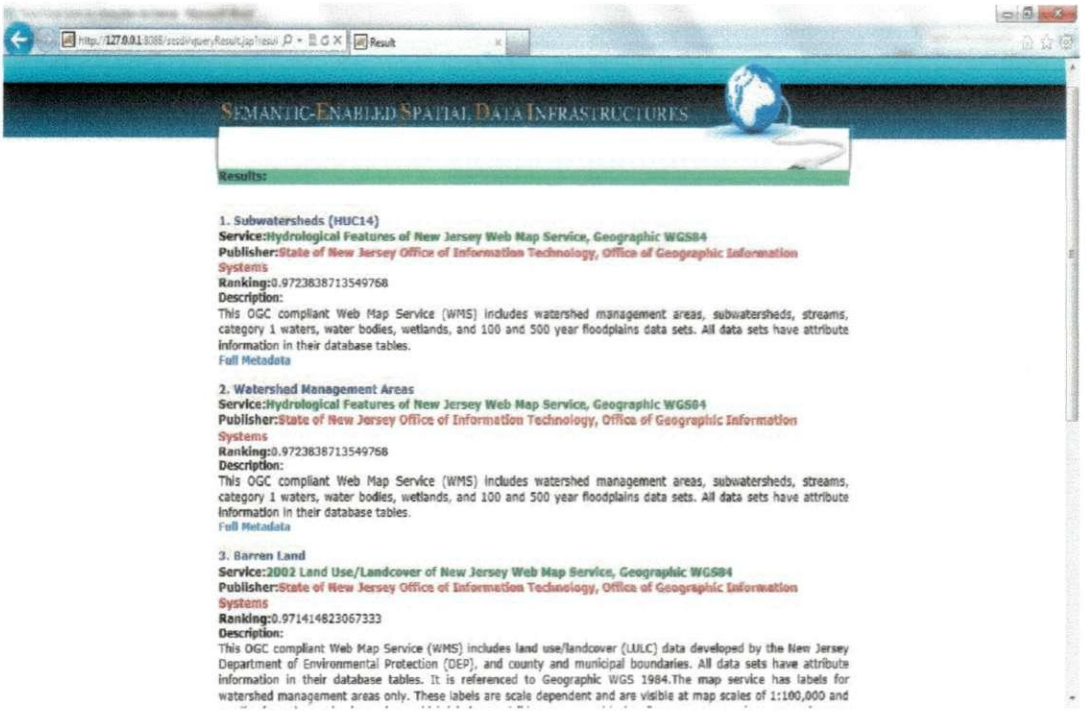

Figura 6.1: Resultado de uma consulta espacial

6.2 *Ranking* temático

Atualmente, muitos trabalhos que usam semântica para melhorar a recuperação de dados geográficos usam soluções baseadas no relacionamento semântico de subsunção, que consiste em recuperar todos os recursos associados a conceitos que são subsumidos pelo conceito definido na consulta do usuário. Este tipo de solução melhora a qualidade das consultas, mas também possui limitações. A principal delas é que não existe distinção entre a relevância dos resultados recuperados durante a consulta. Desta forma, recursos que são associados diretamente ao conceito de busca são julgados com a mesma relevância que recursos associados a conceitos relacionados. Como o número de resultados muitas vezes é grande (sobretudo quando buscas são realizadas com conceitos muito gerais), este tipo de abordagem pode fazer com que muitos resultados

desvantagem deste tipo de abordagem era que a similaridade era determinada através de qualquer número real. Entretanto, o valor de similaridade é mais intuitivo quando expresso através de um valor entre 0 e 1.

Devido às limitações das abordagens consideradas, optou-se pela utilização de uma outra abordagem, conhecida como soma ponderada (do inglês, *weighted sum*). Neste tipo de abordagem, que também é conhecida como análise de decisão com múltiplos objetivos, o valor de um objetivo é calculado como a soma de diversos critérios, no qual para cada critério é atribuído um peso. A vantagem deste método com relação aos demais é que o mesmo permite diferenciar melhor o valor da similaridade a partir dos critérios selecionados, permitindo uma melhor classificação dos resultados recuperados. Além disto, neste método, os pesos permitem que o valor da similaridade seja facilmente ajustado para um valor entre 0 e 1, tornando o valor da métrica mais intuitivo para o usuário.

De acordo com a métrica escolhida, dada uma região geográfica B definida na consulta e um *feature type* T oferecido pela infraestrutura, o *ranking* espacial de T pode ser calculado através da Equação 8. Nesta equação, T_{SPA} representa a região geográfica coberta por T, enquanto S representa o serviço pelo qual T é oferecido.

$$spa\_ranking(T,B) = w_1 * overlap(B,T_{SPA}) + w_2 * relevance(B,S) \quad (8)$$

Onde:

- *spa_ranking* representa a relevância do *feature type* T para a região geográfica requisitada pela consulta;
- *overlap* representa o grau de sobreposição entre a região requisitada e a região coberta pelo *feature type*;
- *relevance* representa a relevância que a região requisitada tem para o serviço pelo qual o *feature type* é oferecido.

A Figura 6.1 mostra os primeiros resultados obtidos através da resolução de uma consulta espacial no SESDI. No cenário desta consulta, o cliente deseja recuperar mapas sobre o estado de *New Jersey*. Este tipo de consulta retorna uma grande quantidade de resultados, uma vez que existem muitos *feature types* cuja extensão geográfica intersecta esta região. A análise desta figura mostra que o SESDI prioriza as camadas

cuja extensão geográfica é mais parecida com a região definida na consulta. Isto acontece devido ao grau de sobreposição. A figura mostra também que os primeiros resultados mostrados são oferecidos por serviços nos quais a região geográfica da consulta possui uma maior relevância. Tal opção ocorre devido ao grau de relevância espacial.

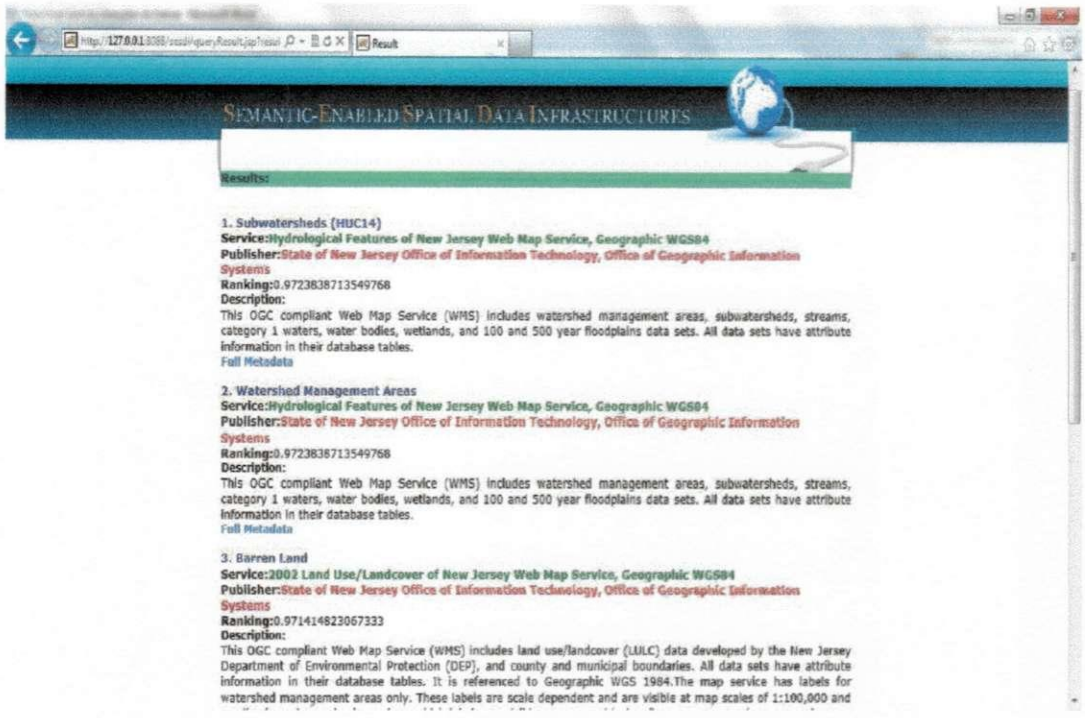

Figura 6.1: Resultado de uma consulta espacial

6.2 *Ranking* temático

Atualmente, muitos trabalhos que usam semântica para melhorar a recuperação de dados geográficos usam soluções baseadas no relacionamento semântico de subsunção, que consiste em recuperar todos os recursos associados a conceitos que são subsumidos pelo conceito definido na consulta do usuário. Este tipo de solução melhora a qualidade das consultas, mas também possui limitações. A principal delas é que não existe distinção entre a relevância dos resultados recuperados durante a consulta. Desta forma, recursos que são associados diretamente ao conceito de busca são julgados com a mesma relevância que recursos associados a conceitos relacionados. Como o número de resultados muitas vezes é grande (sobretudo quando buscas são realizadas com conceitos muito gerais), este tipo de abordagem pode fazer com que muitos resultados

representa o conceito Q, enquanto o nó N_n corresponde ao nó que representa o conceito D.

Para cada caminho identificado entre dois conceitos, pode-se afirmar que existe um conjunto de pesos $W = \{w_1, w_2, w_3, ..., w_{n-1}\}$, no qual cada w_i corresponde ao peso do arco usado para interconectar os nós N_i e N_{i+1}. Os valores destes pesos são usados para determinar o valor do relacionamento semântico entre os dois conceitos. Este valor é determinado através da média aritmética dos pesos que compõem o caminho. Tal escolha se seu porque nesta métrica todos os pesos envolvidos são considerados durante a avaliação deste relacionamento. Desta forma, dado um caminho N entre os conceitos Q e D, o valor do relacionamento semântico (*relationship*) entre os dois conceitos através deste caminho é calculado através da média aritmética dos pesos existentes no conjunto W, como definido na Equação 9:

$$relationship(Q, D, N) = average\{W(N)\} \qquad (9)$$

Onde:

- *relationship* representa o valor do relacionamento semântico entre um conceito de busca (Q) e um conceito (D) usado para a anotação temática do *feature type* que está sendo avaliado;
- W representa o conjunto de pesos associado ao caminho N.

Quando dois conceitos são avaliados, existem situações nas quais existe mais de um caminho entre os mesmos. Para determinar qual dos caminhos disponíveis deve ser considerado para avaliar o valor do relacionamento semântico entre estes conceitos foi introduzida uma métrica chamada de melhor caminho (*bestPath*). De acordo com esta métrica, dado um conjunto de caminhos $P(Q, D) = \{P_1, P_2, ..., P_n\}$, no qual cada P_i, com $1 \leq i \leq n$, é um caminho que liga o conceito Q ao conceito D, o *bestPath*(P) corresponde ao P_i que possui o maior valor para o relacionamento semântico. A escolha deste caminho se deu porque o mesmo preserva os relacionamentos semânticos mais fortes existentes entre os conceitos que estão sendo avaliados.

Depois de definir como o melhor caminho é selecionado, pode-se determinar o valor do relacionamento semântico entre dois conceitos. Dados dois conceitos Q e D, o valor do relacionamento semântico entre estes conceitos é determinado através da

Equação 10. Nesta equação, P(Q, D) corresponde ao conjunto formado por todos os caminhos possíveis que levam do conceito Q para o conceito D.

$$relationship(Q,D) = relationship(Q,D,bestPath(P(Q,D))) \qquad (10)$$

Na Figura 6.3 é destacado, em vermelho, o melhor caminho entre os conceitos *HydrosphereElement* e *River*. Neste caso, o melhor caminho *bestPath*(*HydrosphereElement*, *River*) é uma sequência contendo os seguintes nós, na respectiva ordem {*HydrosphereElement*, *Watercourse*, *River*}. Neste caminho, o valor do relacionamento semântico entre estes conceitos é igual a 0.8, uma vez que os pesos deste caminho são {0.8, 0.8}. A figura também mostra que existe outro caminho possível entre estes conceitos, formado pelos nós {*HydrosphereElement*, *Channel*, *River*}. Entretanto, o conjunto de pesos deste caminho é igual a {0.8, 0.6}, o que faz com que, através do mesmo, o valor do relacionamento semântico entre os conceitos seja igual a 0.7.

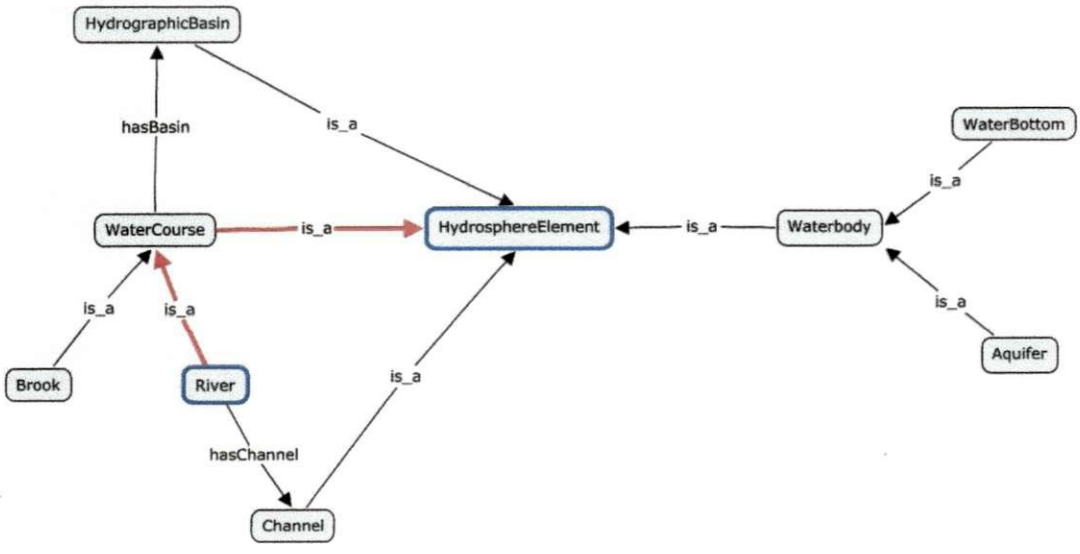

Figura 6.3: Caminho entre dois conceitos na ontologia

A segunda métrica usada para computar a similaridade entre os conceitos é a distância existente entre os mesmos. O seu valor é obtido através do comprimento do melhor caminho que liga os dois conceitos que estão sendo comparados. O uso da distância para avaliar a similaridade entre conceitos definidos foi introduzido por Rada et al. (1989). O objetivo desta variável é garantir que pares de conceitos mais próximos

possuam um grau de similaridade maior do que pares de conceitos que estejam mais distantes, desde que os dois pares tenham o mesmo tipo de relacionamento semântico. Além disto, ela permite garantir o requisito referente ao grau de generalização. Esta medida é inversamente proporcional ao grau de similaridade, ou seja, à medida que a distância entre os conceitos aumenta, a similaridade entre os mesmos diminui.

Uma vez identificados, os valores do relacionamento semântico e da distância são combinados para determinar o grau de similaridade entre dois conceitos. Dados um conceito de busca Q e um conceito D associado ao *feature type* que está sendo avaliado, de forma que Q e D são distintos, a similaridade entre Q e D é calculada através da Equação 11. Caso os dois conceitos sejam idênticos, considera-se que o grau de similaridade entre os mesmos é igual a 1.

$$similarity(Q,D) = w_1 * relationship(Q,D) + w_2 * \left(\frac{1}{dis\tan ce(Q,D)} \right) \qquad (11)$$

Onde:

- *similarity* representa o grau de similaridade entre o conceito de busca Q e o conceito D usado para a anotação do *feature type* que está sendo avaliado;
- *relationship* representa o valor do relacionamento semântico existente entre os conceitos Q e D;
- *distance* representa a distância entre os conceitos Q e D;
- w_1 e w_2 representam os pesos que cada uma destas medidas tem para o cálculo da similaridade. Cada peso deve ter um valor entre 0 e 1, e a soma dos mesmos deve ser sempre igual a 1. Os valores 0.8 e 0.2 são usados, respectivamente, para os pesos w_1 e w_2.

A Tabela 6.3 mostra uma matriz de similaridade que descreve o grau de similaridade entre alguns conceitos mostrados na ontologia da Figura 6.3. As linhas representam os conceitos definidos na requisição do usuário, enquanto as colunas representam os conceitos usados para a anotação dos *feature types*. A análise da tabela mostra que o valor da similaridade entre os conceitos é assimétrico. Por exemplo, se o usuário está procurando pelo tema *HydrosphereElement* e o *feature type* está anotado

com o conceito *WaterCourse*, a similaridade é de 0.84. Contudo, se o usuário procura pelo tema *WaterCourse* e o *feature type* está anotado com o conceito *HydrosphereElement*, o valor da similaridade cai para 0.68, uma vez que nem todos os dados do *feature type* são relevantes para o usuário.

Outra característica importante que pode ser notada é que, à medida em que a profundidade do conceito aumenta, o nível de similaridade diminui. Por exemplo, se o usuário procura pelo tema *HydrosphereElement* e um *feature type* está anotado com o conceito *WaterCourse*, a similaridade possui valor de 0.84. Por sua vez, se outra camada é anotada com o conceito *River* que é ainda mais específico, o valor da similaridade diminui para 0.74. A tabela também mostra que, quando os conceitos são idênticos, a similaridade entre os mesmos é sempre igual a 1.

Tabela 6.3: Matriz de similaridade de conceitos

	Hydrosphere Element	WaterCourse	River	Channel
Hydrosphere Element	1	0.84	0.74	0.84
WaterCourse	0.68	1	0.84	0.66
River	0.58	0.68	1	0.72
Channel	0.68	0.58	0.72	1

6.2.3 O grau de relevância temática

Assim como no *ranking* espacial, o grau de relevância que o tema da consulta do usuário tem para o serviço que oferece um *feature type* que está sendo avaliado é considerado durante o processo de avaliação do *ranking* temático. O objetivo desta medida é avaliar o quanto este tema é relevante para o serviço que o oferece, de forma a permitir que *feature type*s oferecidos por serviços que possuem um maior grau de especialidade no tema requisitado sejam mostrados primeiro para o usuário durante a exibição do resultado da consulta.

O grau de relevância de um tema C para um serviço S é calculado através do mesmo método usado para determinar a relevância espacial (seção 6.1.3). Entretanto, neste caso, a frequência do tema da consulta é determinada através do grau de similaridade entre o conceito definido na requisição e os conceitos usados para a anotação semântica de cada camada oferecida pelo serviço. A sua frequencia inversa,

por sua vez, é determinada através do número de serviços que possuem pelo menos uma camada associada ao conceito de busca ou a um conceito subsumido pelo mesmo. Desta forma, o grau de relevância temático é calculado através da Equação 12:

$$relevance(C,S) = \frac{\sum_{i=1}^{n} similarity(C, Ti_{TEM})}{n} * \log\frac{N}{n_j} \quad (12)$$

6.2.4 Calculando o *ranking* temático

O valor final do *ranking* temático é determinado através da combinação dos valores obtidos para o grau de similaridade entre os conceitos que representam o tema da consulta e o tema usado para a anotação temática do *feature type* e o grau de relevância do tema da consulta para o serviço que o oferece, de acordo com a Equação 13. Assim como nas consultas espaciais (seção 6.1.4), o valor deste *ranking* é determinado através de uma soma ponderada, que usa, respectivamente, os pesos 0.84 e 0.16 para estas métricas.

$$tem\_ranking(T,C) = w_1 * overlap(C, T_{TEM}) + w_2 * relevance(C,S) \quad (13)$$

A Figura 6.4 mostra as primeiras camadas recuperadas por uma consulta temática. No cenário desta consulta, o cliente procura por mapas sobre corpos hídricos, que são representados pelo conceito "*BodyOfWater*". A análise desta figura permite observar que todas as camadas que aparecem entre os primeiros resultados se referem ao conceito definido na requisição. Nesta consulta, são recuperadas também as camadas associadas a outros conceitos, tais como "*River*", "*Lake*" e "*Stream*", que são relacionados ao conceito de busca. Entretanto, devido ao grau de similaridade entre conceitos, estes *feature types* acabam sendo apresentados depois para o cliente. A análise do resultado obtido mostra que as primeiras camadas mostradas para o usuário são oferecidas por um serviço de feições hidrológicas e por um serviço de hidrografia. Isto acontece porque nestes serviços o tema da consulta do usuário possui uma relevância maior do que para outros serviços recuperados.

Figura 6.4: Resultado de uma consulta temática

6.3 *Ranking* temporal

O *ranking* temporal é uma métrica desenvolvida para a resolução de consultas com restrições temporais. Para isto, esta medida avalia o quanto um *feature type* oferecido pela IDE é relevante para a consulta do usuário, considerando apenas a dimensão temporal de ambos. Durante o desenvolvimento do *ranking* temporal, alguns trabalhos existentes envolvendo motores de busca temporais voltados para a recuperação de documentos foram avaliados. Alguns dos trabalhos abordados exploram o uso de semântica para inferir relacionamentos semânticos entre intervalos temporais (ALLEN, 1998) (HUBNER; VISSER, 2003) (VÖGELE et al., 2003). Entretanto, este tipo de abordagem não oferece *ranking*. Por outro lado, alguns motores de busca temporal oferecem *ranking* (ALONSO et al., 2006) (JIN et al., 2010) (MANICA et al., 2010) (STRÖTGEN; GERTZ, 2010) (ALONSO et al., 2011), mas o mesmo é calculado apenas com base no texto dos documentos, de forma que as restrições temporais são usadas apenas durante a seleção dos documentos. Devido a estas limitações, para a implementação do SESDI, optou-se pelo desenvolvimento de uma nova medida de *ranking* temporal.

Durante a resolução de consultas temporais, o SESDI seleciona todos os *feature types* cuja extensão temporal intersecta o intervalo definido na requisição do cliente. Para cada camada selecionada, o *ranking* temporal é determinado a partir de duas

variáveis: o grau de sobreposição e o grau de relevância temporal. Antes de apresentá-las, esta seção discute os requisitos que foram considerados para o desenvolvimento desta métrica.

6.3.1 Os requisitos para o *ranking* temporal

Os requisitos levantados para a métrica que seria usada para a resolução de consultas temporais foram:

- **Informação temporal em dois níveis**: quando um serviço é publicado no serviço de catálogo da IDE, as suas informações temporais podem ser incluídas no seu registro de metadados. Geralmente, quando cadastradas, as informações temporais descrevem a extensão temporal do serviço como um todo, o que limita a qualidade do processo de recuperação de informação. Para superar esta limitação, foi definida a utilização de dois tipos de descrição de informação temporal: um para a descrição do serviço e outro para a descrição de cada *feature type*. Esta característica foi definida para permitir uma descrição mais detalhada acerca da extensão temporal dos serviços, possibilitando melhorar a qualidade das consultas;
- **Extensão temporal representada como intervalo:** informação temporal é geralmente representada de duas formas: pontos no tempo ou intervalos no tempo. Na abordagem baseada em pontos, a informação temporal é representada como um ponto na reta que representa o tempo. Por sua vez, a abordagem baseada em intervalos representa esta informação através de intervalos, que correspondem a um conjunto de pontos composto por pelo menos dois pontos distintos. Para o desenvolvimento do SESDI, foi definido que as extensões temporais dos serviços e *feature types* deveriam ser representadas como um intervalo. Esta opção se deu porque este tipo de representação permite uma maior expressividade para descrever informação temporal. Para ilustrar este tipo de situação, vamos supor um *feature type* que descreve o processo de ocupação de uma determinada região no período de 2005 a 2011. Este tipo de situação, que é bastante comum em serviços de dados geográficos, é difícil de ser expressa através de um único ponto. Entretanto, a mesma situação

pode ser facilmente descrita através do intervalo [2005, 2011]. Ademais, extensões temporais encontradas no serviço de catálogo são geralmente descritas na forma de intervalos;

- **Representação de intervalos persistentes:** uma característica importante de intervalos temporais é que os mesmos possuem um limite inicial e um limite final, que definem, respectivamente, o seu início e o seu término. Entretanto, os limites destes intervalos nem sempre são bem definidos. Isto pode acontecer, por exemplo, quando o valor de um destes limites é desconhecido ou não pode ser precisamente determinado. Outra situação bastante comum acontece quando um limite é irrestrito. Por exemplo, vamos considerar um *feature type* que fornece informações diárias sobre a temperatura da superfície do mar desde 2008. Neste caso, o intervalo referente à sua extensão temporal tem um limite inicial bem definido (2008), mas não possui um limite final. Este tipo de limite é chamado de persistente, e pode ocorrer tanto no limite inferior quanto no limite superior. Como a ocorrência deste tipo de intervalo é bastante comum no oferecimento de dados geográficos, foi definido que o SESDI deveria ser capaz de representar estes tipos de intervalo, bem como processá-los de forma conveniente durante a resolução de consultas;
- **Assimetria:** seja t_1 o intervalo temporal definido na consulta do usuário e t_2 o intervalo temporal associado ao *feature type* que está sendo avaliado. Foi definido que o *ranking* temporal entre estes intervalos deveria depender da ordem em que os mesmos fossem comparados. Para compreender este requisito, vamos considerar uma situação na qual estes intervalos são diferentes e t_1 está contido em t_2. Caso t_1 seja o intervalo solicitado na consulta e t_2 o intervalo coberto pelo *feature type*, a consulta do usuário é completamente resolvida. Entretanto, na situação inversa, a consulta do usuário não é completamente resolvida, o que requer que no segundo caso o valor do *ranking* temporal entre estes intervalos seja menor do que o valor obtido no primeiro caso. Além disto, foi definido que a similaridade entre dois intervalos temporais idênticos deveria ser sempre igual a 1,

enquanto que a similaridade entre dois intervalos disjuntos deveria ser sempre igual a 0;

- **Complemento dos intervalos deveria ser considerado:** foi definido também que a avaliação do *ranking* entre dois intervalos temporais deveria levar em consideração tanto os subintervalos que os mesmos possuem em comum quanto os subintervalos que os mesmos não possuem em comum. O objetivo deste requisito era priorizar *feature types* associados a intervalos temporais mais parecidos com o intervalo definido na consulta. Esta medida também deveria ser inversamente proporcional ao valor do *ranking*. Desta forma, quanto maior fosse o intervalo que os intervalos não tivessem em comum, menor deveria ser o grau de similaridade entre os mesmos.

6.3.2 O grau de sobreposição temporal

A primeira medida usada para avaliar o *ranking* temporal é chamada de grau de sobreposição, e avalia a similaridade entre o intervalo temporal requisitado pela consulta e o intervalo temporal coberto pelo *feature type* que está sendo avaliado. Assim como no *ranking* espacial, o grau de sobreposição temporal é calculado através da equação de Tversky. Desta forma, considerando-se t_1 o intervalo temporal requisitado na consulta do usuário e t_2 o intervalo associado ao *feature type* que está sendo avaliado, o grau de sobreposição entre os mesmos é calculado através da Equação 14.

$$overlap(t_1, t_2) = \frac{length(t_1 \cap t_2)}{length(t_1 \cap t_2) + \alpha * length(t_1 - t_2) + (1-\alpha) * length(t_2 - t_1)} \quad (14)$$

Onde:

- *length* $(t_1 \cap t_2)$ representa a duração, em milissegundos, do intervalo que corresponde à interseção entre os intervalos t_1 e t_2;
- *length* $(t_1 - t_2)$ representa a duração do intervalo que pertence ao intervalo t_1, mas não pertence ao intervalo t_2;
- *length* $(t_2 - t_1)$ representa a duração do intervalo que pertence ao intervalo t_2, mas não pertence ao intervalo t_1;

- A constante α representa o peso que o complemento do intervalo t_1 tem para a avaliação da sobreposição entre os intervalos, enquanto o valor 1 – α representa o peso do complemento de t_2 para esta avaliação. De forma a manter o requisito de assimetria, o valor de α deve ser maior do que 0.5. O valor 0.9 foi definido para esta constante.

A Tabela 6.4 mostra alguns exemplos de intervalos temporais, com diferentes granularidades. A Tabela 6.5, por sua vez, mostra uma matriz contendo os valores do grau de sobreposição entre os intervalos mostrados na tabela anterior, calculados a partir da Equação 14. Nesta tabela, os intervalos das linhas representam a restrição temporal definida na consulta, enquanto que os intervalos das colunas representam a extensão temporal referente ao *feature type* que está sendo avaliado. Através da análise da tabela, é possível perceber que o grau de sobreposição satisfaz o requisito de assimetria. Por exemplo, quando o usuário define uma busca pelo intervalo t_1 e um *feature type* cobre o intervalo t_2, o grau de sobreposição é 0.5209, enquanto na situação inversa o grau de sobreposição aumenta para 0.9072. Este tipo de situação acontece porque no segundo caso a consulta é completamente satisfeita, o que não acontece no primeiro caso.

Tabela 6.4: Exemplos de intervalos temporais

Intervalo	Início	Fim
t_1	01-01-2011	31-12-2011
t_2	01-01-2011	30-06-2011
t_3	01-05-2011	31-05-2011
t_4	01-05-2011	20-05-2011

Tabela 6.5: Matriz de valores de sobreposições temporais

	t_1 (ft)	t_2 (ft)	t_3 (ft)	t_4 (ft)
t_1 (q)	1	0.5209	0.0907	0.0576
t_2 (q)	0.9072	1	0.1817	0.1158
t_3 (q)	0.4731	0.6666	1	0.6574
t_4 (q)	0.3551	0.5412	0.9452	1

6.3.3 O grau de relevância temporal

O grau de relevância temporal é usado para avaliar o quanto um intervalo temporal é relevante para um serviço, permitindo a priorização de *feature types*

oferecidos por serviços que sejam mais especializados no intervalo de tempo requisitado na consulta. O valor desta métrica é calculado através do mesmo método usado para determinar o grau de relevância espacial (seção 6.1.3). Neste método, dado um intervalo temporal t definido na consulta do cliente, a frequência deste intervalo é determinada através do seu grau de sobreposição entre t e os intervalos temporais associados a cada camada oferecida pelo serviço. A sua frequência inversa, por sua vez, é calculada através do número de serviços que possuem pelo menos uma camada cuja extensão temporal o cobre totalmente. A Equação 15 mostra a avaliação desta métrica.

$$relevance(t,S) = \frac{\sum_{i=1}^{n} overlap(t, Ti_{TEMP})}{n} * \log \frac{N}{n_j} \qquad (15)$$

6.3.4 Calculando o *ranking* temporal

Uma vez calculados, os valores do grau de sobreposição e da relevância temporal são combinados, através de uma soma ponderada, para determinar o *ranking* temporal de um *feature type*. Para determinar este *ranking* são usados para estas métricas, respectivamente, os pesos 0.84 e 0.16. Este método é idêntico ao usado para determinar os valores dos *rankings* espacial e temático (seção 6.1.4) e é mostrado através da Equação 16.

$$temp\_ranking(T,t) = w_1 * overlap(t, T_{TEMP}) + w_2 * relevance(t,S) \quad (16)$$

A Figura 6.5 mostra as primeiras camadas recuperadas após a realização de uma consulta temporal feita através do SESDI. No cenário desta consulta, o cliente realiza uma busca por mapas referentes ao ano de 1964. A figura mostra que, devido ao grau de sobreposição temporal, são mostradas primeiramente as camadas que se referem exatamente ao período requisitado pelo usuário. Depois destas camadas, é apresentada uma camada cujo intervalo temporal cobre totalmente o intervalo requisitado na consulta, mas que possui um valor de *ranking* menor devido ao baixo grau de sobreposição entre estes intervalos.

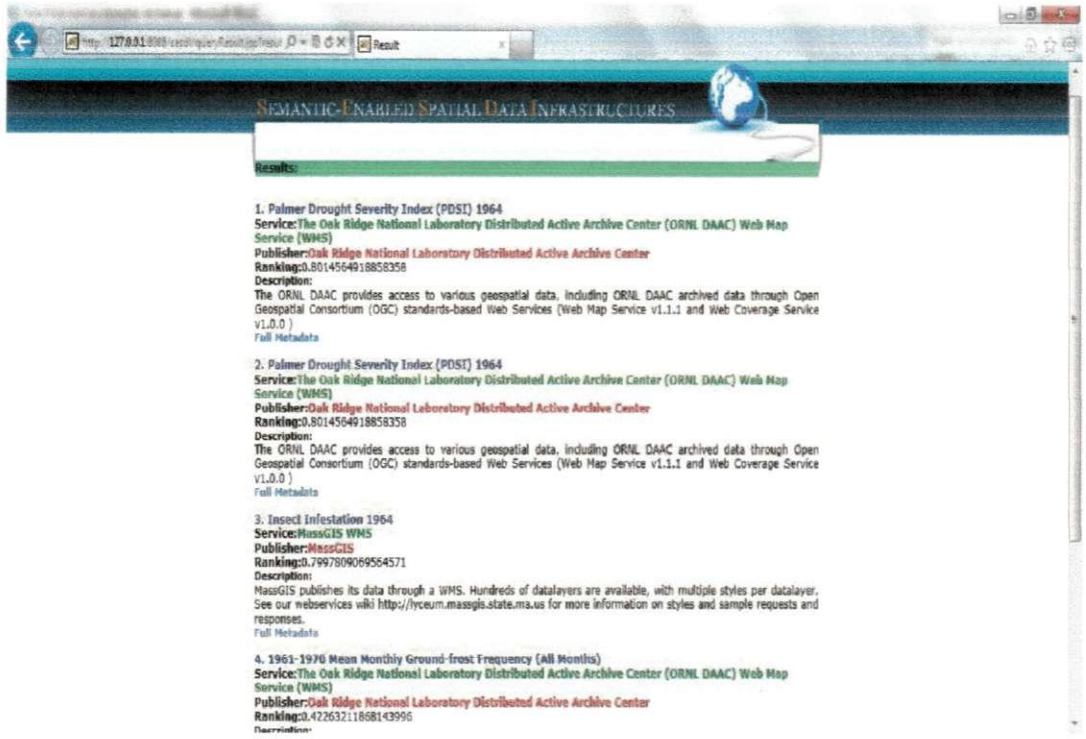

Figura 6.5: Resultado de uma consulta temporal

6.4 *Ranking* global

Os valores obtidos para os *rankings* espacial, temático e temporal permitem avaliar o grau de relevância de cada *feature type* com relação às dimensões espaço, tema e tempo, respectivamente. Depois que estas medidas são avaliadas, elas podem ser combinadas para estabelecer uma medida de similaridade global para a camada. O objetivo desta medida é oferecer um valor de *ranking* global que permita distinguir o quanto cada *feature type* oferecido pela infraestrutura é relevante para a consulta do usuário. Esta medida permite a geração de um *ranking* para organizar os resultados recuperados.

Dados uma consulta Q e um *feature type* T oferecido pela infraestrutura, a similaridade global entre os mesmos pode ser calculada através da Equação 17:

$$rank(Q,T) = w_1 * spaRank(B,T) + w_2 * temRank(C,T) + w_3 * tempRank(t,T) \quad (17)$$

Onde:

- *rank* representa o grau de relevância do *feature type* T para a consulta Q, requisitando a região geográfica B, o conceito C e o intervalo temporal t;

- *spaRank* representa o grau de relevância do *feature type T* para a consulta Q, considerando apenas a dimensão espacial de ambos;
- *temRank* representa o grau de relevância do *feature type T* para a consulta Q, considerando apenas a dimensão temática de ambos;
- *tempRank* representa o grau de relevância do *feature type T* para a consulta Q, considerando apenas a dimensão temporal de ambos;
- w_1, w_2 e w_3 representam os pesos que cada tipo de *ranking* tem para o cálculo do *ranking* global. Diferentemente das outras medidas de *ranking*, os pesos desta equação podem variar de acordo com as consultas. Isto acontece porque uma das dimensões pode ser omitida na requisição. Por exemplo, se uma consulta possui apenas restrições espaciais e temáticas, então é assumido que a soma entre w_1 e w_2 deve ser igual a 1 e que w_3 é 0.

A Tabela 6.6 mostra a configuração dos pesos usados para a resolução de consultas globais. A primeira linha referente aos dados descreve os pesos utilizados quando a consulta possui todos os três tipos de restrição. As demais linhas mostram, respectivamente, os valores dos pesos quando a consulta não possui qualquer restrição temporal, temática e espacial. Entretanto, por questões de flexibilidade, o módulo também permite que o usuário decida qual o peso que deve ser considerado para cada dimensão.

Tabela 6.6: Pesos usados para a resolução de consultas globais

w_1	w_2	w_3
0,08	0,83	0,09
0,09	0,91	0
0,52	0	0,48
0	0,91	0,09

A Figura 6.6 mostra o resultado de uma consulta global. No cenário desta consulta, o cliente está interessado em mapas que mostrem o uso da terra na região de *San Diego* durante o período de 2008. Este tipo de requisição é mais complexo, uma vez que envolve restrição espacial (a região de *San Diego*), temática (uso da terra, representado pelo tema *Land Use*) e temporal (2008). A análise da figura mostra que em todo o banco de dados existem apenas duas camadas que satisfazem todas estas

restrições. A figura mostra também que as duas camadas recuperadas durante a resolução da consulta possuem exatamente o mesmo valor de *ranking*. Isto acontece porque os seus respectivos serviços, embora sejam diferentes, oferecem exatamente as mesmas camadas.

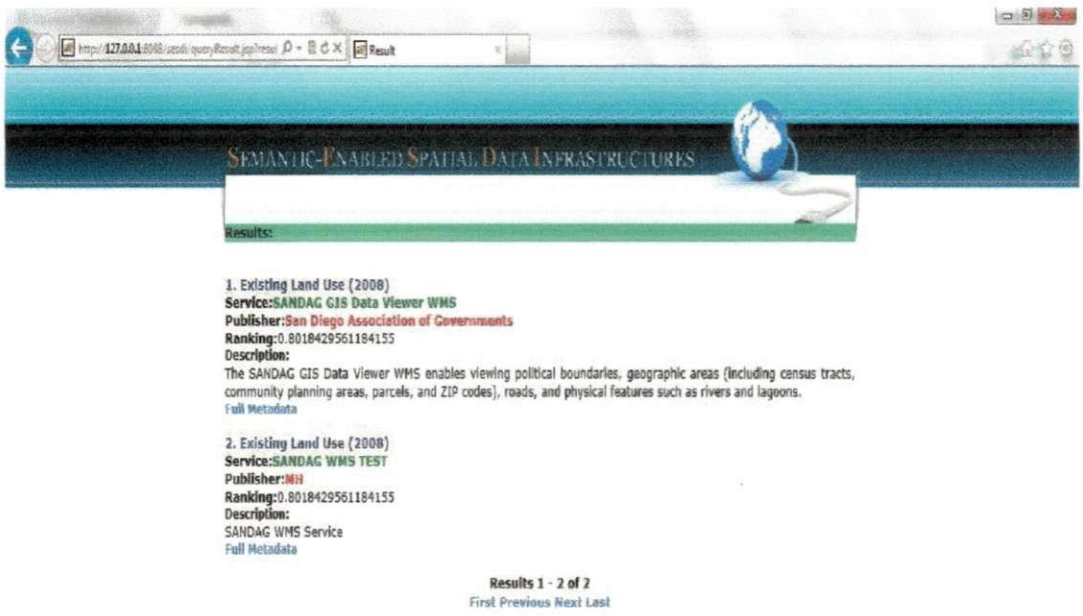

Figura 6.6: Resultado de uma consulta global

6.5 Considerações finais

Este capítulo descreveu o subsistema de recuperação da informação. Ao longo do seu desenvolvimento, o capítulo mostrou como as ideias utilizadas para a recuperação de documentos podem ser adaptadas e reaproveitadas para melhorar a recuperação de dados geográficos. O capítulo discutiu também as métricas que são usadas para determinar os *rankings* espacial, temático, temporal e global. O próximo capítulo discute o processo de validação do arcabouço.

Capítulo 7 – Avaliação Experimental

Este capítulo discute o processo de avaliação experimental do SESDI. Inicialmente, o capítulo descreve o seu processo de prototipação, abordando características como as ferramentas e tecnologias usadas para o seu desenvolvimento, a infraestrutura de dados espaciais utilizada como estudo de caso e as ontologias usadas para a anotação e recuperação dos dados. Depois, o capítulo discute o processo de validação, realizando uma comparação entre os resultados obtidos através do SESDI e os resultados obtidos através do serviço de catálogo.

7.1 Prototipação

O arcabouço proposto por esta tese foi validado através de um protótipo. Este protótipo, que foi totalmente implementado na linguagem de programação Java, foi desenvolvido de acordo com a arquitetura apresentada no Capítulo 4.

O primeiro módulo implementado foi o de gerenciamento de ontologias. Uma importante definição para o seu desenvolvimento, foi a definição das ontologias que seriam usadas durante os processos de anotação temática e de resolução de consultas com este tipo de restrição. Para realizar esta tarefa, foi usada a versão 2.2 da ontologia SWEET[5] (*Semantic Web for Earth and Environmental Terminology*), que é uma ontologia descrita pela NASA[6] (*National Aeronautics and Space Administration*). Esta ontologia é atualmente composta por mais de quatro mil conceitos, que descrevem 197 diferentes domínios de aplicação. Para a escolha desta ontologia, foi considerado que a maior parte das IDEs atuais oferecem dados geográficos referentes a diversos domínios de aplicação. Por isto, a utilização de uma ontologia que descrevesse vários domínios foi fundamental para permitir a anotação e a recuperação de uma maior quantidade de dados. Outra característica que contribui para a utilização desta ontologia é que a mesma já é descrita em OWL, que é a linguagem recomendada pelo W3C para a definição de ontologias. É importante ressaltar, entretanto, que o arcabouço proposto não depende de qualquer ontologia específica, sendo capaz de trabalhar com qualquer

[5] http://sweet.jpl.nasa.gov/ontology/
[6] http://www.nasa.gov/

ontologia que seja adicionada ao seu banco de dados. Para o acesso e o processamento das ontologias usadas pelo SESDI, foi usado o *framework* Jena[7].

O subsistema de coleta de informações foi o segundo módulo implementado. Para a sua implementação, foi definido que a Infraestrutura de Dados Espaciais Norte-americana (NSDI) seria usada como estudo de caso. A escolha desta IDE deu-se devido à grande quantidade de serviços de dados geográficos oferecidos pelo seu serviço de catálogo e pela quantidade de informações contidas em seus registros de metadados. Tais características foram fundamentais para avaliar a qualidade das consultas realizadas pelo SESDI.

Uma vez que a NSDI foi escolhida como estudo de caso, o seu serviço de catálogo foi acessado para a coleta de serviços geográficos. Estes acessos foram realizados usando a API oferecida pela ferramenta *GeoNetwork*[8]. Para cada serviço coletado, o seu documento de funcionalidades foi obtido para a identificação dos seus *feature types*. Este acesso foi realizado usando a API *GeoTools*[9]. Todas as informações obtidas foram processadas, anotadas (espacialmente, semanticamente e temporalmente) e armazenadas no banco de dados do protótipo, que foi implementada usando o sistema de gerência de banco de dados PostgreSQL[10], com o uso da extensão PostGis[11] para o armazenamento, indexação e recuperação de dados geográficos. Esta base de dados conta hoje com 12.914 *feature types*, distribuídos entre 104 serviços.

O último módulo desenvolvido foi o motor de busca. Neste motor, o cliente pode acessar e usar o SESDI através de um conjunto de páginas *web* dinâmicas, que foram implementadas usando a tecnologia *Java Server Pages* e armazenadas em um servidor *Apache Tomcat*.

7.2 Validação

Uma vez implementado, o arcabouço proposto pela tese foi validado. O objetivo desta validação foi comparar os resultados das consultas realizadas através do SESDI com aqueles obtidos através de consultas similares enviadas para o serviço de catálogo da IDE. Durante o processo de avaliação, foram realizados vários experimentos, relativos a quatro tipos de consultas: espacial, temática, temporal e global. Cada

[7] http://jena.apache.org/
[8] http://www.geonetwork-opensource.org/
[9] http://geotools.org/
[10] http://www.postgresql.org.br/
[11] http://postgis.refractions.net/

consulta foi realizada usando tanto o SESDI, através de sua interface gráfica, quanto o serviço de catálogo da NSDI, através da operação *GetRecords*.

Os resultados obtidos para cada abordagem foram comparados através da cobertura e da precisão, que correspondem às principais métricas para a avaliação do desempenho de sistemas de recuperação da informação. A cobertura é obtida através da proporção entre o número de recursos relevantes recuperados em uma consulta e o número de recursos relevantes existentes no sistema. A precisão, por sua vez, é obtida através da proporção entre o número de recursos relevantes recuperados pela consulta e o número total de recursos recuperados pela mesma.

Para que a cobertura e a precisão pudessem ser avaliadas, antes de cada consulta, uma *baseline* foi gerada contendo todas as camadas que eram relevantes para a requisição que estava sendo processada. Estas *baselines* foram determinadas manualmente, através da análise das informações obtidas para os serviços e *feature types* presentes na base de dados do arcabouço. As próximas seções descrevem os resultados obtidos para cada tipo de consulta.

7.3 Avaliação das consultas espaciais

Para validar as consultas espaçais, foram realizadas várias requisições tendo como única restrição a região geográfica de interesse do usuário. Durante este processo, foram realizadas vinte consultas envolvendo localidades de diferentes níveis de abrangência.

Uma importante característica que foi considerada durante a definição das consultas que foram usadas para avaliar as consultas espaciais foi o uso de regiões geográficas referentes a localidades de diferentes abrangências, como local, estadual, regional e nacional. Outro fator considerado durante a escolha destas regiões foi a frequência das mesmas nas camadas dos serviços oferecidos pelo serviço de catálogo da IDE. Para definir as consultas que seriam usadas para a validação, deu-se preferência a localidades que eram cobertas por uma quantidade maior de camadas. Esta característica levava à geração de *baselines* maiores, que permitiam uma melhor avaliação do desempenho das duas abordagens. Este mesmo critério, inclusive, foi considerado durante a validação dos demais tipos de consulta.

As requisições que foram usadas para a validação das consultas espaciais são mostradas na Tabela 7.1. As três primeiras requisições envolvem regiões de nível local. As consultas entre Q4 e Q7, por sua vez, envolvem localidades de nível regional. As

consultas entre Q8 e Q18 envolvem localidades de nível estadual. Finalmente, as consultas Q19 e Q20 requisitam regiões de nível nacional. A Figura 7.1 mostra como a consulta Q1 é realizada através da interface gráfica oferecida pelo SESDI.

Tabela 7.1: Requisições usadas para a validação das consultas espaciais

ID	Consulta Espacial
Q1	Encontre mapas sobre a região de *San Francisco*.
Q2	Encontre mapas sobre a cidade de *San Diego*.
Q3	Encontre mapas sobre a região de *Oak Ridge*.
Q4	Encontre mapas sobre a região de *Washington County*.
Q5	Encontre mapas sobre a região de *Allen County*.
Q6	Encontre mapas sobre a região de *Ashland County*.
Q7	Encontre mapas sobre a região do *Caribe*.
Q8	Encontre mapas sobre o estado de *Rhode Island*.
Q9	Encontre mapas sobre o estado de *Boston*.
Q10	Encontre mapas sobre o estado de *Massachusetts*.
Q11	Encontre mapas sobre o estado de *New York*.
Q12	Encontre mapas sobre o estado de *Maine*.
Q13	Encontre mapas sobre o estado de *Idaho*.
Q14	Encontre mapas sobre o estado de *Ohio*.
Q15	Encontre mapas sobre o estado de *New Jersey*.
Q16	Encontre mapas sobre o estado da *California*.
Q17	Encontre mapas sobre o estado do *Hawaii*.
Q18	Encontre mapas sobre o estado do *Alaska*.
Q19	Encontre mapas sobre os *Estados Unidos*.
Q20	Encontre mapas sobre o *Canadá*.

Os resultados obtidos para cada consulta foram comparados em duas etapas. Na primeira delas, foram analisadas a cobertura e a precisão de cada abordagem com relação ao número de serviços recuperados. O objetivo desta comparação era avaliar a quantidade e a qualidade dos serviços recuperados por cada abordagem. Através desta avaliação, foi possível perceber a quantidade de serviços que possuíam dados relevantes para a consulta, mas que não foram recuperados, caracterizando um problema de cobertura. Da mesma forma, esta observação também possibilitou avaliar a quantidade de serviços que eram recuperados mesmo sem ter dados relevantes para a consulta.

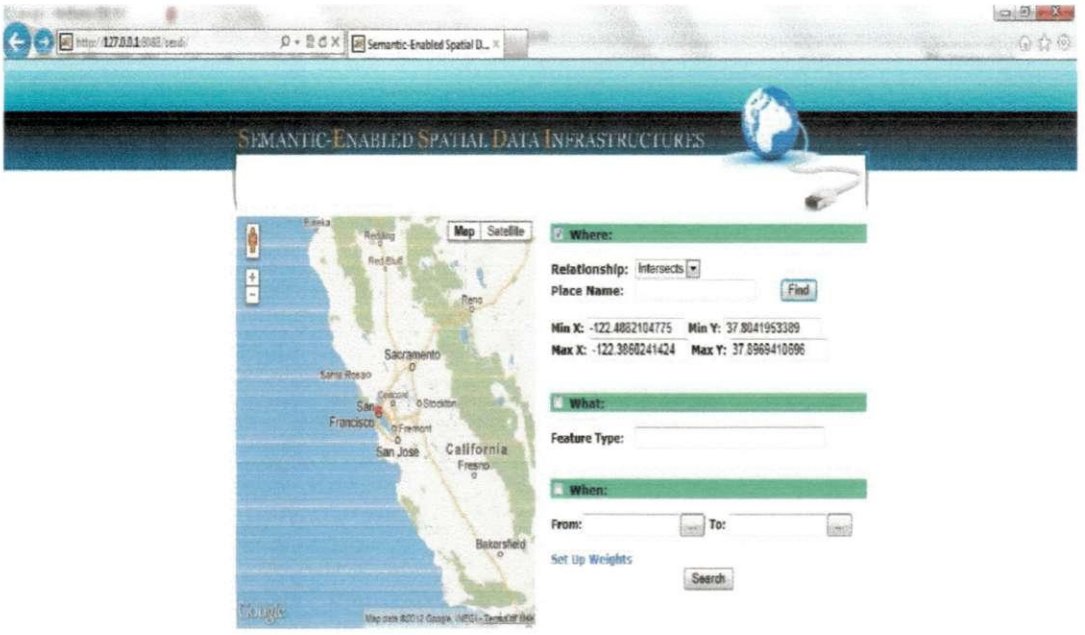

Figura 7.1:Exemplo de uma requisição para uma consulta espacial

Na segunda etapa, as duas abordagens foram comparadas de acordo com o número de *feature types* recuperados por cada abordagem. O objetivo desta comparação era avaliar a quantidade de camadas relevantes que deixavam de ser recuperadas para o cliente devido aos problemas de consistência. Da mesma forma, esta análise também permitiu verificar a quantidade de camadas irrelevantes que foram recuperadas durante cada consulta. É importante ressaltar que os problemas de precisão fazem com que o cliente tenha maior dificuldade para localizar, dentre todos os recursos recuperados pela consulta, aqueles que são de seu interesse.

Para avaliar o desempenho das duas abordagens, para cada consulta realizada, foram criadas duas *baselines*, sendo uma para a comparação em nível de serviços e outra para a comparação em nível de *feature types*. A *baseline* para a comparação em nível de *feature types* continha todas as camadas cujo *bounding-box* intersectava a região geográfica definida na consulta. Por outro lado, a *baseline* usada para a avaliação em nível de serviços continha todos os serviços que ofereciam pelo menos uma camada presente na *baseline* em nível de *feature types*.

A Figura 7.2 mostra os gráficos obtidos durante a validação das consultas espaciais. Nestes gráficos, o eixo y representa o valor obtido para estas métricas, enquanto o eixo x representa as consultas realizadas durante a validação.

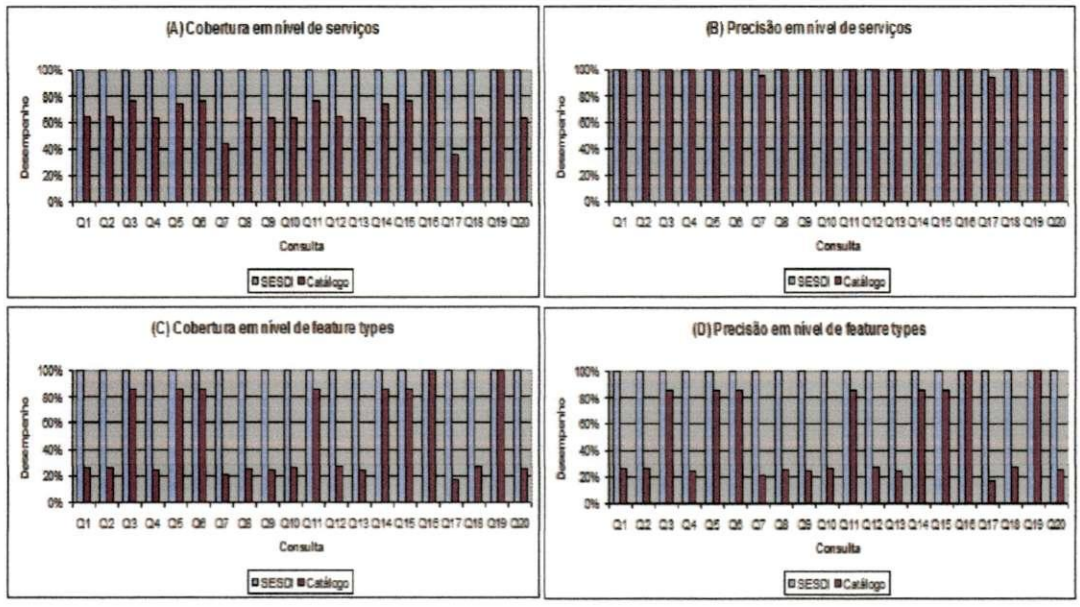

Figura 7.2: Resultados da validação das consultas espaciais

Os gráficos A e B mostram, respectivamente, a comparação dos valores da cobertura e da precisão obtidos a partir de cada abordagem em nível de serviços. Os experimentos realizados mostraram que, com relação à dimensão espacial, o SESDI apresentou melhor cobertura e precisão, conseguindo um desempenho de 100% para ambas as métricas. O serviço de catálogo, por sua vez, obteve uma cobertura média de 66,64% e uma precisão média de 99,29%.

Os valores obtidos para estas métricas podem ser explicados pelo fato de que o SESDI resolve as suas consultas com base nas extensões geográficas oferecidas pelo documento de funcionalidades do serviço, que descreve detalhadamente a região geográfica coberta por cada camada. O serviço de catálogo, por sua vez, resolve consultas espaciais com base na extensão geográfica descrita no registro de metadados do serviço, que usa uma única extensão para representar todas as suas camadas. A análise dos resultados mostrou que a redução da cobertura do catálogo se deve principalmente a problemas de consistência entre a extensão definida no registro de metadados e os valores definidos no documento de funcionalidades. Ademais, no SESDI, todas as informações são recuperadas a partir de um banco de dados geográfico, o que garante este desempenho para a realização de consultas.

Os gráficos C e D da Figura 7.2 mostram, respectivamente, a comparação entre as duas abordagens com relação ao número de *feature types*. Os gráficos mostram que, quando as mesmas são comparadas em nível de camadas, a diferença entre os seus

desempenhos é ainda maior. Neste caso, o SESDI mantém os valores de 100% para a cobertura e para a precisão. Entretanto, a cobertura média do serviço de catálogo cai para 49,34% e a precisão média diminui para 75,60%. A redução da cobertura do serviço de catálogo ocorre porque muitos serviços que deixam de ser recuperados oferecem uma grande quantidade de camadas relevantes. Com relação à precisão, a sua diminuição acontece porque o serviço de catálogo só recupera o serviço como um todo, o que faz com que muitos *feature types* irrelevantes sejam recuperados durante a sua consulta.

7.4 Avaliação das consultas temáticas

A abordagem desenvolvida para a resolução de consultas temáticas foi o segundo tipo de consulta a ser avaliado. Para isto, foram realizadas consultas nas quais a única restrição era o tema requisitado pelo usuário. Na consulta realizada através do SESDI, este tema correspondia a um conceito definido em uma das suas ontologias usadas no estudo de caso. No serviço de catálogo, as consultas temáticas correspondiam a uma busca por palavras-chave, que procuravam por serviços que possuíam os termos da consulta em algum de seus atributos textuais, tais como palavras-chave, título ou descrição.

A Tabela 7.2 descreve as requisições que foram usadas para avaliar o desempenho das consultas temáticas oferecidas pelo SESDI. Para a validação deste tipo de consulta, foram usados temas de diferentes níveis de abstração. Por exemplo, as consultas entre Q1 e Q6 representam requisições por temas que são bastante gerais, sendo representados por classes que possuem uma grande quantidade de subclasses. As demais consultas, por sua vez, requisitam temas mais específicos, que, por consequência, são representados por classes que possuem poucas (ou até nenhuma) subclasses.

A utilização de conceitos de diferentes níveis de abstração foi necessária para avaliar o quanto a utilização de semântica e ontologias poderia melhorar o processo de recuperação de dados geográficos. A Figura 7.3 mostra como a consulta Q1 é realizada através da interface gráfica oferecida pelo SESDI. Nesta figura, o tema requisitado, que é referente à costa, é representado pelo conceito *Coastal Zone*, que é definido pela ontologia de hidrografia usada para a prototipação do SESDI.

Tabela 7.2: Requisições usadas para a validação das consultas temáticas

ID	Consulta Tempatica
Q1	Encontre dados sobre costas.
Q2	Encontre dados sobre hidrologia.
Q3	Encontre dados sobre corpos hídricos.
Q4	Encontre dados sobre geomorfologia.
Q5	Encontre dados sobre áreas urbanas.
Q6	Encontre dados sobre conservação.
Q7	Encontre dados sobre tempestades.
Q8	Encontre dados sobre fronteiras.
Q9	Encontre dados sobre praias.
Q10	Encontre dados sobre a poluição da água.
Q11	Encontre dados sobre bacias de drenagem.
Q12	Encontre dados sobre bacias hidrográficas.
Q13	Encontre dados sobre rios.
Q14	Encontre dados sobre lagos.
Q15	Encontre dados sobre a cobertura do solo.

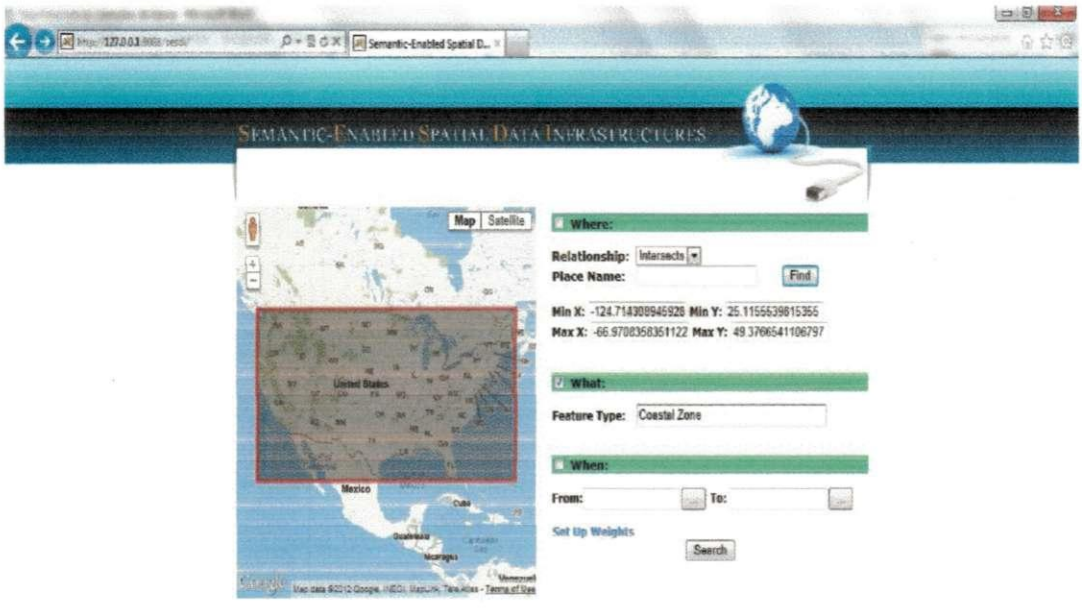

Figura 7.3: Exemplo de uma requisição para uma consulta temática

Durante a validação deste tipo de consulta, as *baselines* de *feature types* continham todas as camadas que ofereciam dados sobre o conceito usado na requisição

ou sobre algum conceito subsumido pelo mesmo. Estas camadas foram identificadas manualmente, através da análise de seus atributos textuais, tais como o seu o título, as suas palavras-chave e a sua descrição textual. Quando as informações disponíveis para o *feature type* eram insuficientes para determinar se o mesmo era ou não relevante para a consulta que estava sendo processada, as informações usadas para a descrição do serviço que o oferecia eram consultadas. Por sua vez, as *baselines* de serviços continham todos os serviços que ofereciam pelo menos um *feature type* relevante para a consulta.

A Figura 7.4 mostra os resultados obtidos durante a validação das consultas temáticas. Os gráficos A e B desta figura mostram, respectivamente, os resultados da avaliação com relação ao número de serviços recuperados. Neste tipo de consulta, o SESDI obteve uma cobertura média de 69,80% e uma precisão média de 75,70%. O serviço de catálogo, por sua vez, obteve uma média de 25,87% e 38,92%, respectivamente, para estas métricas.

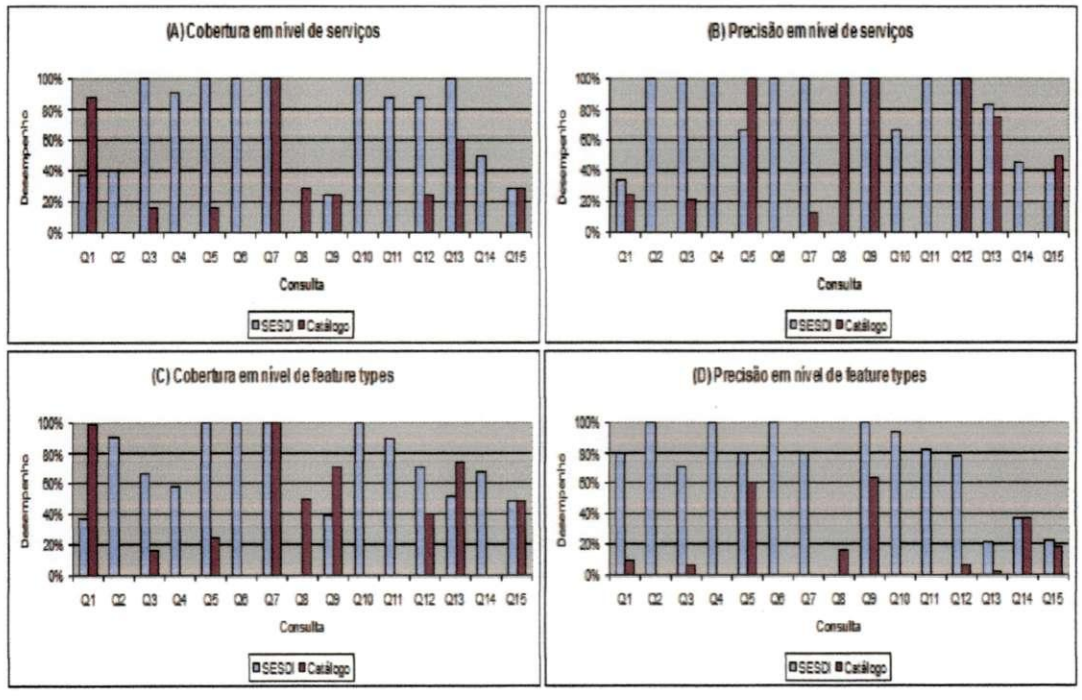

Figura 7.4: Resultados da validação das consultas temáticas

A análise dos resultados obtidos através dos experimentos mostrou que o principal fator que contribui para a grande diferença entre os desempenhos das duas abordagens é o fato de que o serviço de catálogo recupera apenas os serviços que

possuem exatamente o termo da consulta em sua descrição. Em muitos casos, foi possível observar que muitos serviços que possuíam dados relevantes para a consulta deixaram de ser recuperados porque o termo usado para a formulação da consulta não aparece no seu registro de metadados. Esta característica fez com que muitas consultas realizadas pelo catálogo retornem um conjunto vazio.

Outro fato que pôde ser notado durante os experimentos é que a cobertura do das consultas do serviço de catálogo normalmente diminui quando o tema da consulta é muito geral. Esta queda acontece justamente porque o mesmo não consegue identificar os relacionamentos semânticos entre o tema da consulta e as palavras-chave usadas para descrever os serviços no registro de metadados. Desta forma, muitos serviços relevantes, que possuem em sua descrição palavras-chave que são relacionadas ao tema da consulta, acabam não sendo recuperados.

Enquanto o serviço de catálogo tem limitações devido ao uso de palavras-chave, o SESDI, devido ao uso de ontologias, é capaz de recuperar todas as camadas que são associadas a conceitos relacionados ao conceito definido na consulta. Entretanto, devido ao processo de anotação temática automática, algumas anotações relevantes acabam não sendo geradas, prejudicando a cobertura de algumas consultas. Entretanto, foi possível observar que, mesmo com esta limitação, a cobertura média obtida através do SESDI é superior àquela obtida através do serviço de catálogo.

O gráfico B da Figura 7.4 mostra que o SESDI normalmente apresenta uma precisão mais baixa para consultas que envolvem temas mais específicos. Um fator que contribui para esta diminuição é o requisito de assimetria usado para o *ranking* temático. Este requisito permite a recuperação de *feature types* anotados com conceitos que não são subsumidos pelo conceito de busca, o que provoca a recuperação de algumas camadas irrelevantes. Entretanto, esta característica não possui um grande impacto na qualidade da consulta, uma vez que as camadas que se enquadram nesta característica possuem um valor de *ranking* mais baixo e são apresentadas apenas no fim do seu resultado. Ademais, tal característica é proveitosa durante a resolução de consultas que retornam resultados vazios ou muito pequenos para o usuário, uma vez que outros *feature types,* que oferecem dados sobre conceitos relacionados ao conceito de busca, podem ser recomendados para o usuário sem que o mesmo tenha que fazer uma nova requisição.

Assim como na validação do *ranking* espacial, as consultas temáticas também foram comparadas em nível de *feature types*. Os resultados obtidos para a cobertura e a

precisão neste tipo de avaliação são mostrados, respectivamente, nos gráficos C e D da Figura 7.4. Estes gráficos mostram que, neste tipo de avaliação, o desempenho de ambas as abordagens diminui, tanto para a cobertura quanto para a precisão. Entretanto, a diferença entre os seus desempenhos é ainda maior. O SESDI obteve uma cobertura média de 68,17% e uma precisão média de 69,89%. O serviço de catálogo, por sua vez, obteve um desempenho de 34,88% e 14,86%, respectivamente, para estas duas medidas. No serviço de catálogo, estas métricas diminuem devido ao problema de cobertura na recuperação de serviços, que faz com que muitas camadas relevantes deixem de ser recuperados. No SESDI, esta diminuição acontece devido a falhas ocorridas no processo de anotação temática, bem como ao problema de precisão gerado pelo requisito de assimetria.

7.5 Avaliação das consultas temporais

Para a validação das consultas temporais, foram realizadas requisições por diversos intervalos temporais. Para cada consulta realizada através do SESDI, foram requisitados todos os *feature types* cuja extensão temporal fazia interseção com o intervalo de tempo requisitado pelo usuário. Por outro lado, para cada consulta realizada no serviço de catálogo, foram requisitados todos os serviços cuja extensão temporal fazia uma interseção com o intervalo desejado.

A Tabela 7.3 descreve as requisições que foram usadas para avaliar o desempenho das consultas temporais. Para validação deste tipo de consulta, foram usados intervalos temporais de diferentes granularidades. Por exemplo, as consultas entre Q1 e Q8 representam requisições por períodos que representam um ano. Este tipo de granularidade foi mais utilizado porque é identificado com maior frequência nas camadas usadas no estudo de caso. As consultas Q9 e Q10 requisitam dados geográficos referentes a um período de dois anos. As requisições Q11, Q12 e Q13 são referentes a décadas. Finalmente, as consultas Q14 e Q15 requisitam, respectivamente, dados referentes ao período de alguns meses e a um século. A Figura 7.5 mostra como a consulta Q1 é realizada através da interface gráfica oferecida pelo SESDI.

Para validar as consultas temporais, as *baselines* de *feature types* continham todas as camadas cuja extensão temporal intersectava o período de tempo definido na requisição. A extensão temporal de cada camada foi avaliada manualmente, através das informações contidas no seu título e na sua descrição textual. Nos casos nos quais a extensão temporal do *feature type* não podia ser determinada através destes atributos, as

informações temporais do registro de metadados que descreve o seu respectivo serviço foram usadas. A *baseline* de serviços continha todos os serviços que possuíam pelo menos uma camada relevante para a consulta.

Tabela 7.3 : Requisições usadas para a validação das consultas temporais

ID	Consulta Temporal
Q1	Encontre dados referentes ao ano de 1935.
Q2	Encontre dados referentes ao ano de 1964.
Q3	Encontre dados referentes ao ano de 1997.
Q4	Encontre dados referentes ao ano de 1999.
Q5	Encontre dados referentes ao ano de 2000.
Q6	Encontre dados referentes ao ano de 2005.
Q7	Encontre dados referentes ao ano de 2008.
Q8	Encontre dados referentes ao ano de 2009.
Q9	Encontre dados referentes ao período entre 2002 e 2004.
Q10	Encontre dados referentes ao período entre 2006 e 2008.
Q11	Encontre dados referentes à década de 1980.
Q12	Encontre dados referentes à década de 1990.
Q13	Encontre dados referentes à década de 2000.
Q14	Encontre dados referentes ao período a partir de 2012.
Q15	Encontre dados referentes ao período entre 1800 e 1899.

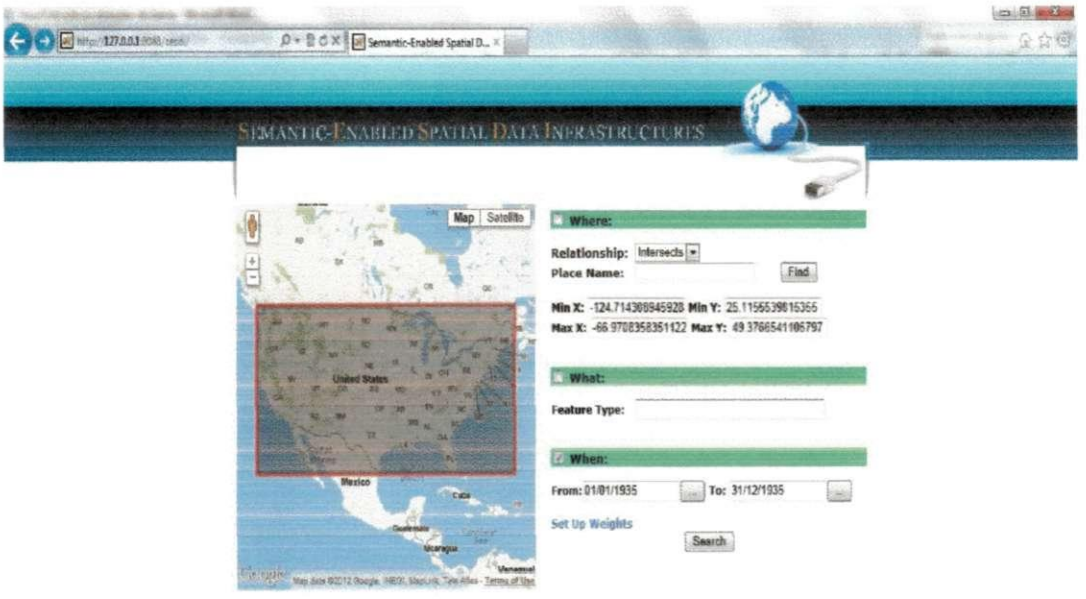

Figura 7.5: Exemplo de uma requisição para uma consulta temporal

Os resultados obtidos durante a validação das consultas temporais são mostrados na Figura 7.6. A análise do gráfico A mostra que o SESDI permitiu um aumento na cobertura deste tipo de consulta. Enquanto o mesmo obteve uma cobertura média de 72,77%, o serviço de catálogo obteve um valor de 54,57% para esta métrica. Esta diferença pode ser explicada pelo fato de que o SESDI resolve consultas temporais com base nas informações obtidas para cada *feature type*, enquanto o serviço de catálogo resolve estas consultas com base nas informações que descrevem o serviço como um todo. Tal característica permite que o SESDI recupere camadas relevantes mesmo que o registro de metadados do seu serviço não ofereça nenhuma informação temporal ou ofereça informações inconsistentes, o que não é possível para o serviço de catálogo.

A análise do gráfico B mostra que, com relação à precisão, as duas abordagens tiveram um desempenho bastante parecido. Os resultados obtidos mostram que o SESDI obteve um melhor desempenho em algumas consultas, enquanto o serviço de catálogo obteve uma precisão melhor em outras requisições. A precisão média do SESDI foi de 88,55%, enquanto que a precisão média do serviço de catálogo foi de 90,03%. A análise dos resultados mostrou que a perda de precisão do SESDI ocorre devido a alguns erros cometidos durante o processo de anotação temporal automática. Este tipo de situação acontece porque algumas expressões temporais não são identificadas durante o processo de anotação ou são interpretadas e anotadas de forma incorreta, o que leva a erros durante o processo de recuperação destes dados.

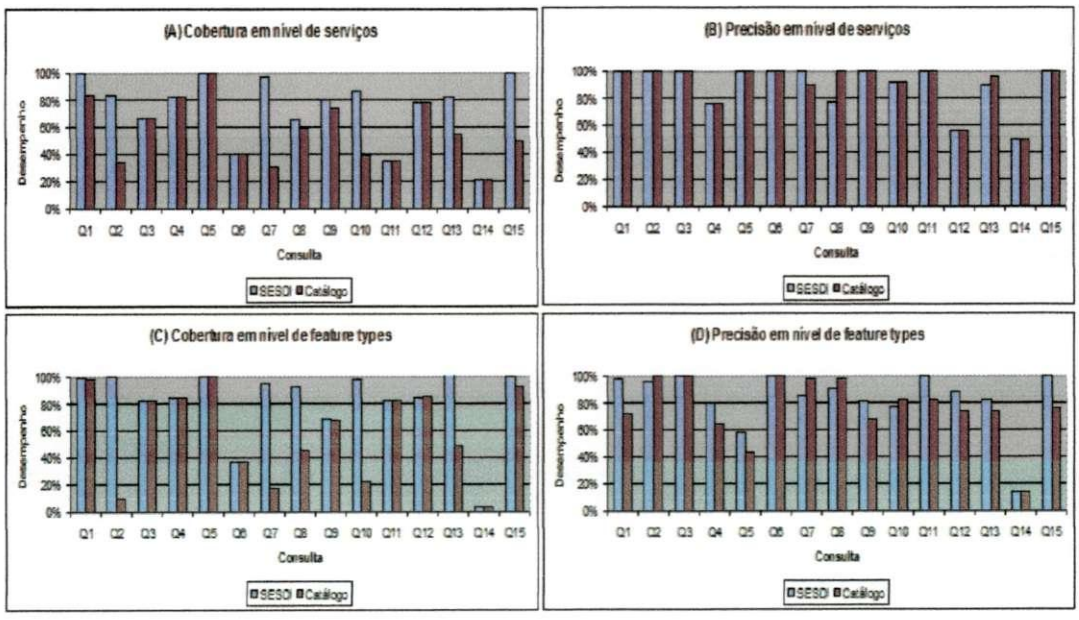

Figura 7.6: Resultados da validação das consultas temporais

O gráfico C mostra a comparação da cobertura das duas abordagens com relação ao número de *feature types* recuperados. A análise deste gráfico mostra que quando a cobertura das duas abordagens é comparada em nível de *feature types*, a diferença do desempenho entre as mesmas é ainda maior. Neste tipo de avaliação, o SESDI apresentou uma cobertura média de 81,90%, enquanto que o serviço de catálogo teve uma cobertura média de 55,97%. As razões que levam a esta diferença são as mesmas que afetam a cobertura em nível de serviços. Entretanto, o grande número de camadas relevantes oferecidas por alguns serviços que não são recuperados pelo serviço de elevam a diferença entre os desempenhos das duas abordagens.

Quando a precisão é comparada em nível de *feature types*, o resultado obtido é bastante parecido com o da comparação através do número de serviços. Esta informação é mostrada no gráfico D. Neste tipo de comparação, o SESDI teve uma precisão média de 83,51%, enquanto que o serviço de catálogo teve uma precisão média de 76,60%. Os fatores que levam a esta diferença são os mesmos que afetam a precisão em nível de serviços.

7.6 Avaliação das consultas globais

Para avaliar o desempenho das consultas globais, foram usadas requisições que tinham restrições espaciais, temáticas e temporais. No SESDI, cada uma destas consultas era composta por uma região geográfica (representada por um *bounding-box*), um tema (representado por um conceito de uma ontologia) e um intervalo temporal. No serviço de catálogo, estas requisições eram compostas por uma região geográfica, uma palavra-chave referente ao tema e um intervalo temporal. Em ambas as abordagens, eram recuperados apenas os recursos que satisfaziam todas as restrições definidas na requisição.

As consultas usadas para validar o desempenho das consultas globais são mostradas na Tabela 7.4. Durante a validação deste tipo de consulta, foram usadas consultas que envolviam a combinação de algumas restrições usadas durante a avaliação das consultas espaciais, temáticas e temporais. A Figura 7.7 mostra como a consulta Q1 é realizada através da interface gráfica oferecida pelo SESDI.

Para validar o desempenho de cada abordagem durante a resolução deste tipo de consulta, a *baseline* de *feature types* contendo todas as camadas que satisfaziam as três restrições definidas na requisição. Estas camadas foram identificadas através da análise das informações contidas no documento de funcionalidades e no registro de metadados

de cada serviço presente no estudo de caso. Para cada requisição, também foi criada uma *baseline* de serviços, contendo todos os serviços que possuíam pelo menos uma camada relevante para a consulta.

Tabela 7.4: Requisições usadas para a validação das consultas globais

ID	Consulta Global
Q1	Encontre dados sobre os corpos hídricos de New Jersey no ano de 2000.
Q2	Encontre dados sobre a geomorfologia do Alaska em 2001.
Q3	Encontre dados sobre a geomorfologia da Califórnia em 2000.
Q4	Encontre dados sobre os rios de Idaho entre 2002 e 2004.
Q5	Encontre dados sobre as áreas urbanas dos Estados em 2002.
Q6	Encontre dados sobre a hidrologia de Boston em 2001.
Q7	Encontre dados sobre a poluição da água em Massachusetts em 1964.
Q8	Encontre dados sobre os lagos de New Jersey entre 2001 e 2003.
Q9	Encontre dados sobre as bacias hidrográficas de Boston em 1964.
Q10	Encontre dados sobre o uso da terra em San Diego no ano de 2008.
Q11	Encontre dados sobre a costa dos Estados Unidos na década de 1980.
Q12	Encontre dados sobre a cobertura do solo na Califórnia durante a década de 1970.
Q13	Encontre dados sobre a cobertura do solo nos Estados Unidos entre 2000 e 2009.
Q14	Encontre dados sobre a costa de New Jersey entre 1900 e 1950.
Q15	Encontre dados sobre áreas de conservação dos Estados Unidos no ano de 2010.

Os resultados obtidos através da validação das consultas globais são mostrados na Figura 7.8. Os gráficos A e B mostram os resultados obtidos pela avaliação em nível de serviços. A análise destes gráficos mostra que houve uma grande diferença entre os desempenhos das duas abordagens. O SESDI teve uma cobertura média de 80,24% e

uma precisão média de 78,89%. O serviço de catálogo, por sua vez, teve uma cobertura média de 21,35% e uma precisão média de 27,78%. Estes valores podem ser explicados pelo fato de que os resultados das consultas globais são afetados por uma combinação dos fatores que influenciam no desempenho das consultas espaciais, temáticas e temporais. Além disto, os baixos índices de cobertura e precisão do serviço de catálogo se devem principalmente às restrições temáticas, uma vez que estes serviços têm grandes limitações para resolver este tipo de consulta.

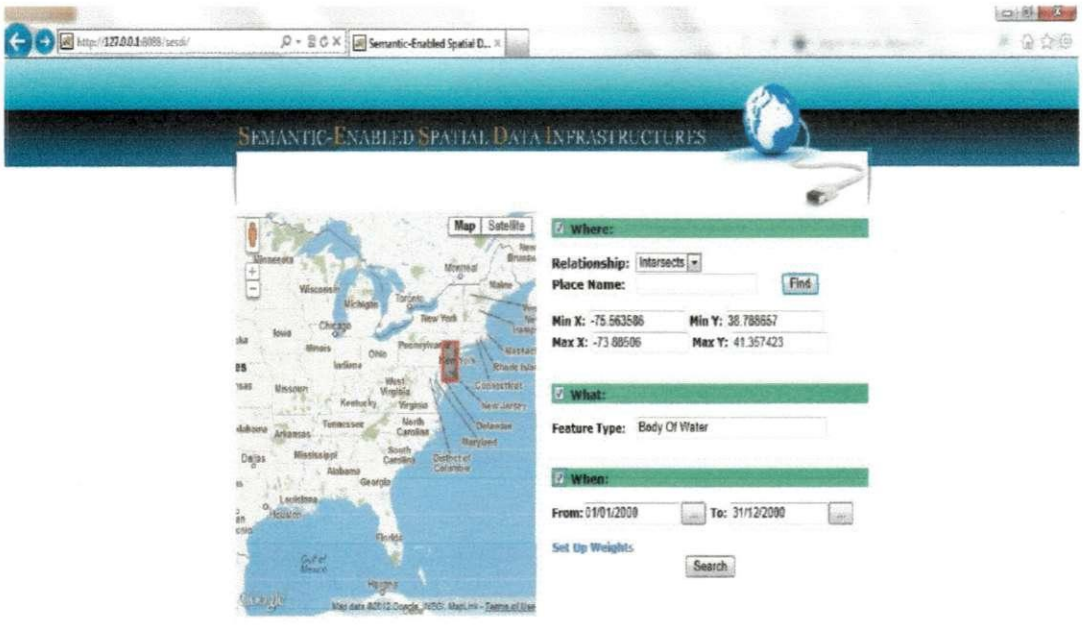

Figura 7.7: Exemplo de uma requisição para uma consulta global

Quando as duas abordagens são comparadas em nível de *feature types*, pode-se perceber que a diferença entre seus desempenhos é ainda maior. Estes valores são mostrados nos gráficos C e D da Figura 7.8. A justificativa para estes valores consiste no fato de que todos os problemas de cobertura e precisão inerentes às dimensões espacial, temática e temporal influenciam nos valores obtidos para estas métricas.

Durante a validação das consultas globais, o SESDI teve uma cobertura média de 72,00% e uma precisão média de 76,17%. Por outro lado, o serviço de catálogo teve uma cobertura média 22,96% e uma precisão média de 8,56%. Assim como nos demais tipos de consulta, a baixa precisão do serviço com relação ao número de camadas recuperadas é justificada pelo fato de que o mesmo só retorna o serviço como um todo, o que faz com que um alto número de camadas relevantes sejam recuperadas.

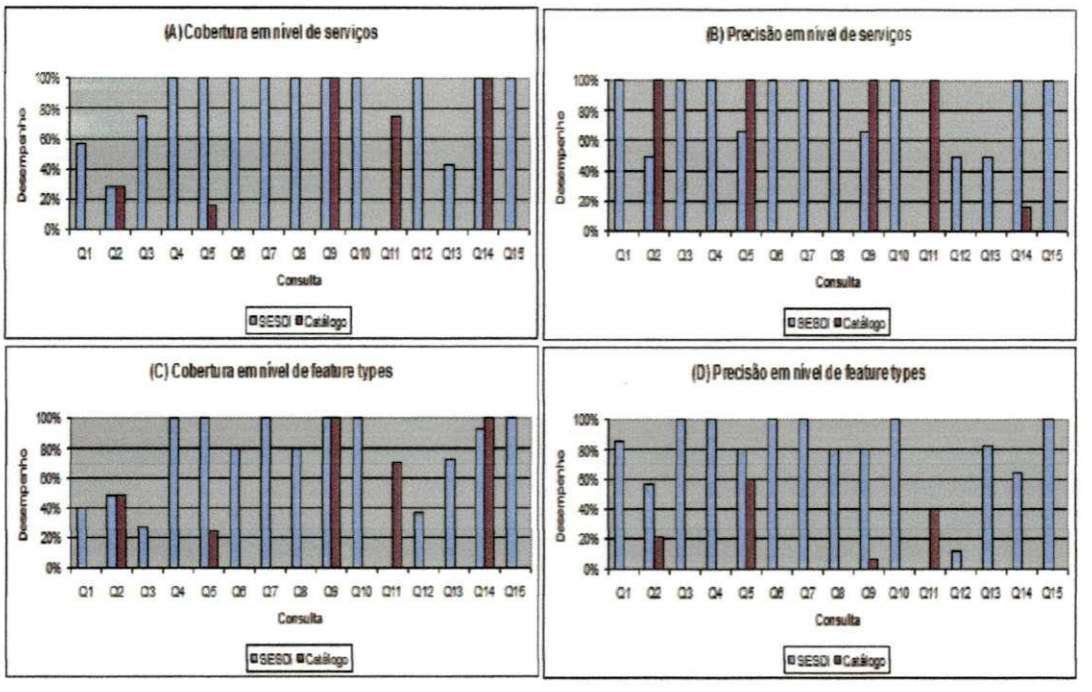

Figura 7.8: Resultados da validação das consultas globais

7.7 Considerações finais

Este capítulo apresentou o processo de avaliação experimental do SESDI. Ao longo do deu desenvolvimento, foram discutidos o estudo de caso usado para a validação. Além disto, o capítulo descreveu o processo de avaliação experimental usado para validar o desempenho do SESDI durante a resolução de consultas espaciais, temáticas, temporais e globais. Tal validação foi feita através da comparação entre os resultados obtidos através do SESDI e aqueles obtidos através do serviço de catálogo atualmente oferecido pela IDE usada como estudo de caso.

A avaliação dos resultados obtidos mostrou que o arcabouço proposto por esta tese melhora consideravelmente a cobertura das consultas para todas as dimensões avaliadas. Os resultados também mostraram que alguns erros cometidos durante o processo de anotação automática reduzem o desempenho do arcabouço em algumas consultas que envolvem restrições temáticas e temporais, principalmente com relação à precisão. Entretanto, foi possível notar que, mesmo com estas limitações, o arcabouço proposto teve um desempenho superior ao do serviço de catálogo da IDE na maior parte das consultas.

Embora os resultados obtidos através da avaliação experimental tenham sido bastante satisfatórios, é importante ressaltar que existem alguns fatores podem influenciar a validade destes resultados. O principal fator é o fato de que atualmente não

existe um *benchmark* disponível para validar a solução proposta pela tese. A inexistência desta ferramenta fez com que todo o processo de avaliação tivesse que ser realizado com base em *baselines* geradas manualmente. O próximo capítulo conclui a tese e mostra os trabalhos futuros desta pesquisa.

Capítulo 8 - Conclusão

Este capítulo apresenta a conclusão deste trabalho de tese. Ao longo do seu desenvolvimento, são discutidas as conclusões que podem ser obtidas a partir de sua avaliação, as contribuições oferecidas pelo desenvolvimento da pesquisa e as publicações obtidas ao longo do trabalho. No final, são discutidos os trabalhos futuros desta pesquisa.

8.1 Conclusões

Nos últimos anos, infraestruturas de dados espaciais têm desempenhado um papel muito importante como a solução para garantir a interoperabilidade de dados geográficos produzidos e oferecidos por diferentes organizações. Tal interoperabilidade é alcançada através da especificação de um conjunto de normas e padrões que devem ser seguidos por todos os seus participantes, enquanto a legislação e a articulação entre as organizações ajudam a definir questões como a custódia e as políticas de acesso e utilização destes dados.

Visando facilitar a recuperação dos dados geográficos disponíveis, as IDEs atuais oferecem um serviço de catálogo, que é usado tanto por seus provedores quanto por seus clientes. Enquanto os provedores de dados geográficos usam este serviço para anunciar os dados que oferecem, clientes o utilizam para localizar os dados de seu interesse. Embora tenham facilitado a recuperação de dados geográficos, os serviços de catálogo atuais ainda possuem limitações que dificultam a realização desta tarefa. A utilização de um único registro para descrever todos os dados oferecidos por um serviço de dados geográficos e o uso de palavras-chave para a resolução de consultas baseadas em tema são algumas das características que prejudicam a cobertura e a precisão de suas consultas. Tais limitações dificultam a localização e a utilização dos dados geográficos que já estão disponíveis, e que poderiam ser usados para diversas finalidades como, por exemplo, o auxílio a processos de tomadas de decisão em diferentes áreas.

Visando superar estas limitações, esta tese propôs SESDI, um arcabouço usado para facilitar a recuperação de dados geográficos oferecidos por uma IDE. Para isto, é utilizado um modelo de dados mais detalhado, que armazena informações acerca das características espaciais, temáticas e temporais de cada *feature type* oferecido por um serviço de dados geográficos. O arcabouço propõe ainda uma série de medidas de

ranking, nas quais as métricas tradicionais da recuperação da informação clássica, voltada para a localização de documentos, são adaptadas e reutilizadas para melhorar a recuperação de dados geográficos. Outra característica importante desta solução é a utilização de ontologias para melhorar a recuperação de dados baseada em tema. Finalmente, é importante ressaltar que todas as informações usadas pelo modelo proposto são identificadas e extraídas automaticamente, a partir das informações contidas no registro de metadados e no documento de funcionalidades do serviço, sem impor qualquer *overhead* aos provedores de dados geográficos.

A avaliação experimental, que comparou os resultados das consultas realizadas através do serviço de catálogo de uma IDE real e já funcional com aqueles obtidos através de consultas realizadas através do arcabouço proposto, mostrou que a solução proposta é viável. Os experimentos realizados mostraram que o SESDI, apesar de uma pequena redução na precisão em alguns tipos de consulta, melhora consideravelmente a cobertura das consultas que envolvem restrições espaciais, temáticas, temporais e globais. Os resultados obtidos a partir da avaliação experimental mostraram que todas as hipóteses levantadas para o desenvolvimento desta pesquisa eram válidas.

8.2 Contribuições

O desenvolvimento desta pesquisa ofereceu contribuições para as comunidades de informação geográfica e recuperação da informação. Estas contribuições foram:

- o desenvolvimento de um modelo de recuperação da informação geográfica para melhorar a recuperação de dados espaciais, que considera e armazena informações em nível de serviço e de *feature types*;
- uma nova abordagem para a anotação semântica de dados geográficos;
- o desenvolvimento de uma medida de *ranking* que permite avaliar a relevância de cada *feature type* oferecido pela IDE para uma consulta do usuário, considerando apenas a dimensão espacial de ambos;
- o desenvolvimento de uma medida de *ranking*, baseada em ontologias, que permite avaliar a relevância de cada *feature type* oferecido pela IDE para uma consulta do usuário, considerando apenas a dimensão temática de ambos;

- o desenvolvimento de uma medida de *ranking* que permite avaliar a relevância de cada *feature type* oferecido pela IDE para uma consulta do usuário, considerando apenas a dimensão temporal de ambos; e
- o desenvolvimento de uma medida de *ranking* que permite avaliar a relevância de cada *feature type* oferecido pela IDE para uma consulta do usuário, a partir de restrições impostas para duas ou três dimensões.

8.3 Resultados obtidos

Ao longo do desenvolvimento desta pesquisa, foram obtidas seis publicações. Estas publicações foram feitas em diferentes tipos de veículos, como:

- periódico nacional (ANDRADE; BAPTISTA, 2011) e internacional (ANDRADE et al., 2011);
- capítulos de livros internacionais (ANDRADE et al., 2011) (ANDRADE; BAPTISTA, 2012);
- artigos completos em conferência nacional (ANDRADE; BAPTISTA, 2010) e internacional (ANDRADE et al., 2012) (este trabalho foi indicado para concorrer ao prêmio de melhor artigo da conferência ICEIS 2012);

8.4 Trabalhos futuros

Alguns trabalhos ainda podem ser desenvolvidos para continuar a pesquisa relativa a esta tese. Estes trabalhos podem incorporar novas características ao arcabouço proposto, visando superar algumas das limitações existentes na versão atual. Os principais trabalhos que podem ser realizados são:

- **Integração do SESDI a um serviço de catálogo:** um importante trabalho futuro a ser desenvolvido consiste em integrar o SESDI à ferramenta *GeoNetwork*. Esta ferramenta, que é um *software* livre e de código aberto, é atualmente usada por muitas infraestruturas de dados espaciais para a implementação de seus serviços de catálogo. Consequentemente, tal integração, além de permitir a utilização do arcabouço em várias IDE, vai permitir que as suas funcionalidades sejam acessadas por outras aplicações de software;

- **Utilização da *DBpedia*:** outro trabalho futuro consiste em utilizar a *DBpedia*[12] durante o processo de anotação temática de *feature types*. Algumas das características fornecidas por esta biblioteca, como o uso de ontologias para descrever a semântica dos recursos oferecidos pela *Wikipedia* e a ligação destes recursos a objetos existentes em outras bases de dados podem melhorar a anotação temática e, consequentemente, o desempenho do SESDI durante a resolução deste tipo de consulta;
- **Melhorar a interface gráfica:** um trabalho futuro necessário é a melhoria da interface gráfica. Tal melhoria vai permitir o desenvolvimento de um portal geográfico, no qual o SESDI poderá ser acessada e utilizada livremente por qualquer usuário com acesso à Internet;
- **Validação dos algoritmos de *ranking*:** embora o desempenho do arcabouço proposto já tenha sido validado com relação à cobertura e à precisão de suas consultas, a validação dos resultados obtidos a partir dos algoritmos de *ranking* ainda é necessária. Uma estratégia que pode ser utilizada para realizar esta validação é a realização de uma avaliação junto ao usuário, de forma a verificar se os primeiros resultados recuperados em cada consulta realmente correspondem à informação requisitada pelo usuário;
- **Melhorar a avaliação da similaridade entre conceitos:** atualmente, a avaliação da similiradade entre conceitos definidos em uma ontologia é baseada apenas no relacionamento semântico e na distância entre ambos. Uma forma de melhorar esta avaliação consiste em considerar a altura destes conceitos. A avaliação desta variável é importante porque a mesma fornece informações sobre o nível de abstração de cada conceito e o nível de detalhes com o qual cada conceito é descrito;
- **Recuperação de serviços de processamento de dados geográficos:** a versão atual do SESDI somente permite a recuperação de serviços de dados geográficos. Um importante trabalho futuro a ser desenvolvido é a implementação de um motor de busca para a recuperação de serviços de

[12] http://dbpedia.org

processamento de dados geográficos, oferecidos através da plataforma de serviços *web* ou do padrão WPS (*Web Processing Service*) (OGC, 2007b). Diferentemente da versão atual do SESDI, que foca nos dados oferecidos pelos serviços, este motor de busca deverá focar nas funcionalidades oferecidas por cada serviço e na descrição da semântica tanto dos dados e condições necessárias para a execução do serviço quanto das informações que são produzidas após a sua execução. A descoberta automática de composições de serviços, nas quais dois ou mais serviços são combinados para resolver uma solicitação do usuário, também é um importante trabalho a ser implementado;

- **Recuperação em nível de feições:** outro trabalho futuro que pode ser implementado se refere à localização de informações em nível de feições. Este tipo de consulta é voltado apenas para camadas oferecidas por serviços WFS e consiste em identificar, dentre todas as *features* oferecidas por uma camada, apenas aquelas que satisfazem os critérios de busca definidos pelo usuário. O uso de semântica e ontologias para resolver este tipo de problema de forma eficiente e escalável ainda representa um grande desafio para as comunidades de banco de dados e de sistemas de informação geográfica;
- **Incorporação de outros tipos de restrição:** a versão atual do SESDI melhora a recuperação de *feature types* geográficos em consultas baseadas em todas as três dimensões inerentes a dados geográficos. Entretanto, novos tipos de filtro podem ser implementados para melhorar estas buscas. Exemplos de novos filtros que podem ser implementados incluem a recuperação de dados com restrição de escala ou resolução, a busca por informações de proveniência e confiabilidade dos dados e a busca por tipos de mapas específicos, como a localização de mapas de curvas de nível. Entretanto, o maior desafio para o desenvolvimento destes novos filtros consiste em identificar automaticamente as informações necessárias para esta implementação, sem gerar *overhead* para provedores no momento de documentação dos dados; e
- **Utilização de outro estudo de caso:** no momento em que o tema desta tese foi definido, o objetivo inicial era que o arcabouço proposto fosse aplicado e utilizado para a recuperação de dados geográficos oferecidos

pela Infraestrutura Nacional de Dados Espaciais Brasileira (INDE). Entretanto, a necessidade de validação da solução da proposta, através de um grande número de serviços de dados geográficos e de metadados descritos de forma mais detalhada, somada às restrições de tempo para a realização da pesquisa, fez com que outra IDE tivesse que ser usada como estudo de caso durante a pesquisa. Com isto, um importante trabalho futuro seria aplicar o arcabouço desenvolvido nesta tese à INDE.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLEN, J. F. Maintaining Knowledge about Temporal Intervals. **Communications of the ACM**, v. 26, n. 11, p. 832–843, 1984.

ALONSO, G.; et. al. **Web services: concepts, architectures and applications**. Berlim: Springer Verlag, 2004.

ALONSO, O.; GERTZ, M.; BAEZA-YATES, R. A. Clustering of search results using temporal attributes. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 29, 2006, Seattle, Washington, USA. **Proceedings**...New York, ACM Press, 2006 p. 597-598.

ALONSO, O. et al. Temporal Information Retrieval: Challenges and Opportunities. In: INTERNATIONAL TEMPORAL WEB ANALYTICS WORKSHOP, 1, 2011, Hyderabad, India. Proceedings... Aachen, CEUR, 2011 p. 1-8.

ANDRADE, F. G.; BAPTISTA, C. S. Using semantic similarity to improve information discovery in spatial data infrastructures. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA, 11, 2010, Campos do Jordão, São Paulo. **Anais...** São José dos Campos: MCT/INPE, 2010. p. 45-56.

ANDRADE, F. G.; BAPTISTA, C. S. Using Semantic Similarity to Improve Information Discovery in Spatial Data Infrastructures. **Journal of Information and Data Management**, v. 2, n. 2, p.181-194, 2011.

ANDRADE, F. G., BAPTISTA, C. S. An Ontology-based Approach to Support Information Discovery in Spatial Data Infrastructures. In: RUCKERMAN, C. (Ed.). **Integrated Information and Computing Systems for Natural, Spatial, and Social Sciences**. 1.ed. Hershey: IGI Global, 2012. (forthcoming)

ANDRADE, F. G.; BAPTISTA, C. S.; SCHIEL, U. A Temporal Search Engine to Improve Geographic Data Retrieval in Spatial Data Infrastructures. In: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS (ICEIS), 14, 2012, Wroclaw, Poland. (forthcoming)

ANDRADE, F. G.; LEITE JR, F. L.; BAPTISTA, C. S. Using Distributed Semantic Catalogs for Information Discovery on Spatial Data Infrastructures. In: ZHAO, P.; DI,

L. (Eds.) **Geospatial Web Services: Advances in Information Interoperability**. 1. ed. Hershey: IGI Global, 2011.

ANDRADE, F. G.; LEITE JR, F. L.; BAPTISTA, C. S. Using Federated Catalogs to Improve Semantic Integration among Spatial Data Infrastructures. **Transactions in GIS**, v. 15, n. 5, p. 707-722, 2011.

ANKOLEKAR, A.; et al. DAML-S: Semantic markup for web services. In: SEMANTIC WEB WORKING SYMPOSIUM, 1, 2001, Stanford, CA. **Proceedings**... Stanford: IOS Press, 2002. p. 411-430.

ANZLIC Spatial Data Infrastructure Committee. **Implementating the Australian Spatial Data Infrastructure**. Disponível em: <http://www.anzlic.org.au>. Acesso em: 11 janeiro 2011.

ATHANASIS, N.; et al. Towards semantics-based approach in the development of geographic portals. **Computers & Geosciences** , v. 35, n. 2, p. 301-308, 2009.

BAADER, F.; et al. **The Description Logic Handbook: Theory, Implementation, Applications**. Cambridge: Cambridge University Press, 2003.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. London: Addison Wesley, 1999.

BATCHELLER, J. K.; REITSMA, F. Implementing feature level semantics for spatial data discovery: supporting the reuse of legacy data using open source components. **Computers, Environment and Urban Systems**, v. 34, n. 1, p. 333–344, 2010.

BERNARD, L.; et al. The European geoportal - one step towards the establishment of a European Spatial Data Infrastructure. **Computers, Environment and Urban Systems**, v. 29, n. 1, p. 15-31, 2005.

BERNARD, L.; et al. Ontology-Based Discovery and Retrieval of Geographic Information in Spatial Data Infrastructures. **International Journal of Geographical Information Science**, v. 20, n. 3, p. 233-260, 2006.

BERNARD, L.; CRAGLIA, M. SDI - From Spatial Data Infrastructure to Service Driven Infrastructure. In: RESEARCH WORKSHOP ON CROSS-LEARNING ON SPATIAL DATA INFRASTRUCTURES AND INFORMATION INFRASTRUCTURES, 1, Twente, Netherlands. **Proceedings**...Twente: ITC, 2005.

BERNERS-LEE, T; HENDLER, J; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34- 43, 2001.
BIANCHINI, D.; et al. Ontology-based methodology for e-service discovery. **Information Systems**, n. 31, p. 361-380, 2006.

BOWERS, S.; LUDÄSCHER, B. An ontology-driven framework for data transformation in scientific workflows. In: INTERNATIONAL WORKSHOP ON DATA INTEGRATION IN THE LIFE SCIENCES, 1, 2004, Leipzig, Alemanha. **Proceedings...** Berlin: Springer, 2004 p. 1-16.

BRUIJN, J.; et al. **Web Service Modelling Ontology (WSMO)**. 2004. Disponível em: <http://www.w3.org/Submission/WSMO/>. Acesso em: 11 janeiro 2011.

CHEN, N.; et al. A capability matching and ontology reasoning method for high precision OGC web service discovery. **International Journal of Digital Earth**, v. 4, n. 6, p. 449-470, 2011.

Comissão Nacional de Cartografia. **Plano de ação para implantação da INDE – Infraestrutura Nacional de Dados Espaciais**. 2010. Disponível em: <http://www.concar.ibge.gov.br/arquivo/PlanoDeAcaoINDE.pdf>. Acesso em: 11 janeiro 2011.

Comitê de Estruturação de Metadados Geoespaciais. **Perfil de metadados geoespaciais do Brasil**. Disponível em: <http://www.sieg.go.gov.br/downloads/Perfil_de_Metadados.pdf>. 2009. Acesso em: 11 janeiro 2011.

DALTIO, J.; MEDEIROS, C. B. **Aondê: um serviço web de ontologias para interoperabilidade em sistemas de biodiversidade**. Dissertação (Mestrado Acadêmico em Ciência da Computação) – Universidade de Campinas, Campinas, 2008.

DARPA Agent Markup Language. Disponível em: <http://www.daml.org/about.html>. Acesso em: 11 janeiro 2011.

Dublin Core. **Using Dublin Core**. Disponível em: <http://dublincore.org/documents/usageguide/>. Acesso em: 11 janeiro 2011.

EGENHOFER, M. Towards the geospatial semantic web. In: ACM INTERNATIONAL SYMPOSIUM ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 1, 2002, MacLean, VA, USA. **Anais....**Nova York, NY: ACM Press, 2002 p. 1- 4.

European Umbrella Organisation for Geographic Information. **La CLEF Project**. 1999. Disponível em: <http://eurogi.org/115-gi-resources/eurogi-ec-projects/107-laclef.html>. Acesso em: 11 janeiro 2011.

ERL, T. **Service-Oriented Architecture: Concepts, Technology and Design**. New Jersey: Prentice Hall, 2005.

Federal Geographic Data Committee. **The National Spatial Data Infrastructure.** 2005. Disponível em: <http://www.fgdc.gov/library/factsheets/documents/nsdi.pdf>. Acesso em: 11 janeiro 2011.

Federal Geographic Data Committee - Metadata Ad Hoc Working Group. **Content Standard for Digital Geospatial Metadata**. 1998. Disponível em: <http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf>. Acesso em: 11 janeiro 2011.

FENSEL, D.; et al. OIL: An Ontology Infrastructure for the Semantic Web. **IEEE Intelligent Systems**, v. 16, n. 2, p. 38-45, 2001.

FENSEL, D. et al. OIL and DAML+OIL: Ontology Languages for the Semantic Web. In: DAVIES, J; FENSEL, D.; VAN HARMELEN, F (Ed.). **Towards The Semantic Web: Ontology-driven Knowledge Management**. 1. ed. New Work: Willey, 2003.

FIELDING, R. T. **Architectural Styles and the Design of Network-based Software Architectures**. Dissertação (Ph.D. Information and Computer Science) – University of California, 2000.

FOX, E. A.; SHAW, J. A. Combination of multiple searches. In: TEXT RETRIEVAL CONFERENCE, 2, 1993, Maryland, USA. **Proceedings**...Maryland: National Institute of Standards and Technology, 1993, p. 243-252.

GE J.; QIU Y. Concept Similarity Matching Based on Semantic Distance. In: INTERNATIONAL CONFERENCE ON SEMANTICS, KNOWLEDGE AND GRID, 4, 2008, Beijing, China. **Proceedings**...Washington: IEEE Computer Society, 2008 p. 380-383.

Global Spatial Data Infrastructure. **The SDI Cookbook**. 2004. Disponível em: <http://www.gsdidocs.org/GSDIWiki/index.php?title=Main_Page>. Acesso em: 11 janeiro. 2011.

GUARINO, N. Formal ontology, conceptual analysis and knowledge representation. **International Journal of Human Computer Studies**, v. 43, n. 5-6, p. 625-640, 1995.

HOLLAND, P.; et al. The Global Spatial Data Infrastructure initiative and its relationship to the vision of a Digital Earth. In: INTERNATIONAL SYMPOSIUM ON DIGITAL EARTH, 1, 1999, Beijing, China. **Proceedings**...Marrickville: Science Press, 1999.

HÜBNER, S; VISSER, U. **Temporal Representation and Reasoning for the Semantic Web**. Disponível em: <http://www.tzi.de/fileadmin/resources/publikationen/tzi_berichte/TZI-Bericht-Nr._28.pdf>. Acesso em: 27 julho 2012.

International Organization for Standardization. **Geographic Information – Temporal Schema**. 2002. Disponível em: <http://www.iso.org/iso/catalogue_detail.htm?csnumber=26013>. Acesso em: 11 janeiro 2011.

International Organization for Standardization. **Geographic Information – Metadata**. 2003. Disponível em: <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber= 26020>. Acesso em 11 janeiro 2011.

International Organization for Standardization. **Data elements and interchange formats - Information interchange - Representation of dates and times**. Disponível em: <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020>. Acesso em 11 janeiro 2011.

International Organization for Standardization. **Geographic Information – Rules for Application Schema**. 2005. Disponível em: <http://www.iso.org/iso/catalogue_detail.htm?csnumber=39891>. Acesso em: 11 janeiro 2011.

International Organization for Standardization. **Geographic Information – Metadata – XML Schema Implementation**. 2007. Disponível em: <http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32557>. Acesso em 11 janeiro 2011.

JANOWICZ K. Sim-DL: towards a semantic similarity measurement theory for the Description Logic ALCNR in geographic information retrieval. In: ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS, 1, 2006, Montpellier, France. **Proceedings...** Berlin: Springer, 2006, p. 1681-1692.

JANOWICZ, K.; WILKES, M.; LUTZ, M. Similarity-based information retrieval and its role within spatial data infrastructures. In: INTERNATIONAL CONFERENCE ON GEOGRAPHIC INFORMATION SCIENCE, 5, 2008, Park City, USA. **Proceedings**... Berlin: Springer-Verlag, 2008, p. 151-167.

JIN, P. et al. CT-Rank: A Time-aware Ranking Algorithm for Web Search. **Journal of Convergence Information Technology**, v. 5, n. 6, p. 99-111, 2010.

KLIEN, E.; LUTZ, M.; KUHN, W. Ontology-Based Discovery of Geographic Information Services - An Application in Disaster Management. **Computers, Environment and Urban Systems**, v. 30, n. 1, p. 102-123, 2006.

KUHN, W. *Geospatial semantics: why, of what, and how?*. In SPACCAPIETRA, S.; ZIMÁNYI, E. (Ed.). **Journal of Data Semantics**. 3. ed. Berlin: Springer, 2005, p. 1-24.

LEMMENS, R.; et al. Enhancing geo-service chaining through deep service descriptions. **Transactions in GIS**, v. 6, n. 1, p. 849-871, 2007.

LI, W.; et al. Semantic-based web service discovery and chaining for building an Arctic spatial data infrastructure. **Computers & Geosciences**, v. 37, n. 1, p. 1752-1762, 2011.

LUTZ, M. Ontology-based descriptions for semantic discovery and composition of geoprocessing services. **Geoinformatica**, v. 11, n. 1, p. 1-36, 2007.

LUTZ, M.; KLIEN, E. Ontology-based Retrieval of Geographic Information. **International Journal of Geographical Information Science**, v. 20, n.3, p. 233-260, 2006.

LUTZ, M.; KOLAS, D. Rule-Based Discovery in Spatial Data Infrastructure. **Transactions in GIS**, v. 11. n.3, p. 317-336, 2007.

MACÁRIO, C. G. N.; SOUSA, S. R.; MEDEIROS, C. B. Annotating geospatial data based on its semantics. In: ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 1, 2009, Seattle, USA. **Proceedings**...New York, NY: ACM Presss, 2009 p. 81-90.

MAGUIRE, D. J.; LONGLEY, P. A. The Emergence of Geoportals and their Role in Spatial Data Infrastructures. **Computers, Environment and Urban Systems**, v. 29, n. 1, p. 3–14, 2005.

MANICA, E.; DORNELES, C. F.; GALANTE, R. M. Supporting Temporal Queries on XML Keyword Search Engines. **Journal of Information and Data Management**, v. 1, n. 3, p. 471-486, 2010.

MARTIN, D.; et al. Bringing Semantics to Web Services: The OWL-S Approach. In: CARDOSO, J; SHETH, A. (Ed.), **Semantic Web Services and Web Process Composition**. 1. ed. Berlin: Springer, p. 26-42, 2005.

MOHAMMADI, H.; et al. Geo-Web Service Tool for Spatial Data Integrability. In: AGILE CONFERENCE ON GI SCIENCE, 11, 2008, Girona, Spain **Anais...**Berlin: Springer, 2008 p. 401-413.

Open Geo Spatial Consortium. **OGC Web Map Service Interface**. 2004. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=4756>. Acesso em: 11 janeiro 2011.

Open Geospatial Consortium. **CSW-ebRIM Registry Service - Part 1: ebRIM profile of CSW**. 2005. Disponível em: <http://portal.opengeospatial.org /files/index.php?artifact_id=31137>. Acesso em: 11 janeiro 2011.

Open Geospatial Consortium. **OpenGIS Filter Encoding Implementation Specification**. 2005. Disponível em: <http://portal.opengeospatial.org /files/index.php?artifact_id=31137>. Acesso em: 11 janeiro 2011.

Open Geospatial Consortium. **Web Feature Service Implementation Specification**. 2005. Disponível em: <https://portal.opengeospatial.org/files/?artifact_id=8339>. Acesso em: 11 janeiro 2011.

Open Geospatial Consortium. **OpenGIS Catalogue Services Specification**. 2007. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=21460>. Acesso em: 11 janeiro 2011.

Open Geospatial Consortium. **OpenGIS Web Processing Service.** 2007. Disponível em: <http://portal.opengeospatial.org/files/index.php?artifact_id=24151>. Acesso em 11 janeiro 2011.

PUSTEJOVSKY J. et al. TimeML: Robust Specification of Event and Temporal Expressions in Text. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL SEMANTICS, 5, 2003, Tilburg, Netherlands. **Proceedings**...Tilburg, ACL/SIGSEM, 2003 p. 28-34.