

ODILON CORRÊA DA SILVA

**USO DE TESAURO E *TOPIC MAPS* NA RECUPERAÇÃO DE
METADADOS EM INFRAESTRUTURAS DE DADOS ESPACIAIS:
UM ESTUDO DE CASO DO ECOSSISTEMA ANTÁRTICO**

Dissertação apresentada à
Universidade Federal de Viçosa, como
parte das exigências do Programa de Pós-
Graduação em Ciência da Computação,
para obtenção do título de *Magister
Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2009

ODILON CORRÊA DA SILVA

**USO DE TESAURO E *TOPIC MAPS* NA RECUPERAÇÃO DE
METADADOS EM INFRAESTRUTURAS DE DADOS ESPACIAIS:
UM ESTUDO DE CASO DO ECOSISTEMA ANTÁRTICO**

Dissertação apresentada à
Universidade Federal de Viçosa, como
parte das exigências do Programa de Pós-
Graduação em Ciência da Computação,
para obtenção do título de *Magister
Scientiae*.

APROVADA: 17 de dezembro de 2009

José Luis Braga
(Co-orientador)

Elpídio Inácio Fernandes Filho

Marco Antônio Casanova

Alcione de Paiva Oliveira

Jugurta Lisboa Filho
(Orientador)

*Dedico esta conquista a meus pais, que tiveram
a sabedoria de entender que um precioso
bem a ser construído para seus filhos
é o conhecimento. A vocês, toda
minha gratidão e meu amor.
Obrigado papai e mamãe!*

AGRADECIMENTOS

A Deus, pela força, saúde e determinação para realizar este trabalho. Agradeço também por colocar as pessoas certas em meu caminho, que sempre me auxiliaram, não só nesse trabalho, como em toda a vida.

A meus pais, Helvécio e Maria das Graças, e irmãos, José Faustino e Maria Auxiliadora, por todo amor, apoio e solidariedade durante esta jornada. Agradeço por sempre acreditarem em mim. Sem eles, nada disso seria possível.

À Gabriela, “meu anjo”, pelo amor, pela paciência, pela dedicação e pelo incentivo constante. Ela que também faz parte dessa conquista. A todos os familiares: tios e primos, pelo grande incentivo.

Ao professor e orientador Jugurta Lisboa Filho, em especial, pelo convívio amigo, companheirismo, diálogo, equilíbrio entre elogios e críticas, pelo exemplo de pessoa e profissional que é. Agradeço por acreditar em mim.

Aos meus co-orientadores, professor José Luis Braga e professora Karla Borges, pelos conselhos e pelas contribuições na identificação e correção de falhas durante o desenvolvimento deste trabalho.

A todos do corpo docente do Departamento de Informática da UFV e ao Altino Souza Filho, secretário da Pós-Graduação, pela pronta ajuda perante as burocracias do mestrado.

A todos os colegas, do mestrado e da graduação, que me incentivaram, ajudaram e contribuíram direta ou indiretamente para a realização desse trabalho, em especial a Bruno Satler, Clayton Fraga, Cristóferon Bueno e Thiago Miranda, companheiros de trabalho e estudo. Ao Ronoel Júnior e Vitor Fontes, alunos do Curso de Ciência da Computação da UFV, pela parceria e pelas contribuições no desenvolvimento desta pesquisa.

Aos meus colegas de trabalho e a toda a equipe da Cientec, em especial ao Paulo Márcio e Alessandro de Freitas.

Ao professor Elpídio Inácio, um dos idealizadores do projeto TERRANTAR, e seu orientado Everton Poelking, que contribuíram para o enriquecimento desta pesquisa.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerias – FAPEMIG e ao Conselho Nacional de Pesquisa – CNPq / MCT / CT-Info, pelo financiamento parcial do projeto.

BIOGRAFIA

ODILON CORRÊA DA SILVA é filho de Helvécio Ferreira da Silva e Maria das Graças Corrêa da Silva, brasileiro, nascido em 18 de outubro de 1978, na cidade de Coronel Fabriciano, no estado de Minas Gerais.

No ano de 2000, após concluir o ensino médio no Colégio Técnico de Coronel Fabriciano, ingressou no curso de graduação em Sistemas de Informação, no Centro Universitário do Leste de Minas Gerais, concluído no ano de 2004.

Em 2007, iniciou o mestrado em Ciência da Computação, na Universidade Federal de Viçosa – UFV, defendendo sua dissertação em dezembro de 2009.

Trabalhou como programador e analista de sistemas no Centro Universitário do Leste de Minas Gerais de agosto de 1999 a março de 2006. Atualmente, é analista de sistemas na Cientec – Consultoria e Desenvolvimento de Sistemas, desde abril de 2006.

SUMÁRIO

LISTA DE FIGURAS.....	vii
LISTA DE TABELAS.....	viii
LISTA DE ABREVIATURAS.....	ix
RESUMO.....	x
ABSTRACT	xi
1 INTRODUÇÃO	1
1.1 O problema e sua importância	2
1.2 Objetivos.....	3
1.3 Organização da dissertação	4
2 REFERENCIAL TEÓRICO	5
2.1 Infraestrutura de dados espaciais	5
2.2 Representação do conhecimento	8
2.2.1 Tesouro.....	9
2.2.2 Ontologia	12
2.2.3 <i>Topic maps</i>	14
2.3 Recuperação da informação	16
2.4 Trabalhos correlatos.....	18
3 UMA ARQUITETURA PARA SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO EM CATÁLOGOS DE METADADOS	22
3.1 Visão geral	22
3.2 Componentes do sistema	24
3.3 Módulo Geração da Base Indexada	26
3.4 Módulo Avaliação da Base Indexada.....	27
3.5 Módulo Consulta Base Indexada.....	28
3.6 Estrutura da arquitetura	29
4 ASPECTOS DE IMPLEMENTAÇÃO.....	31
4.1 Visão geral	31
4.2 Criação e carga do repositório	32
4.3 Funcionalidades do protótipo	35
5 ESTUDO DE CASO	41
5.1 TERRANTAR - Ecossistemas terrestres da antártica.....	41
5.2 Infraestrutura de dados espaciais do TERRANTAR	42
5.2.1 Ferramentas computacionais	42
5.2.2 Funcionalidades e limitações.....	44
5.3 Criação e evolução da base semântica.....	45
5.4 Testes do protótipo.....	47
5.4.1 Teste I: consulta com termos	51
5.4.2 Teste II: consulta com termos relacionados.....	52
5.4.3 Teste III: consulta com termos e especializações	52
5.4.4 Comparação e avaliação dos testes do protótipo	53
6 CONCLUSÕES.....	54

REFERÊNCIAS BIBLIOGRÁFICAS	56
APÊNDICE A	61
A.1 Elementos do Tesauro – ABC do TERRANTAR	61

LISTA DE FIGURAS

Figura 2.1: Estrutura de IDE orientada a serviços.....	6
Figura 2.2: Modelo para representação de um tesouro.....	10
Figura 2.3: Elementos da norma <i>topic maps</i>	15
Figura 2.4: Modelo alternativo para representação de um tesouro.....	19
Figura 2.5: Navegação em mapas dirigida por ontologia.....	20
Figura 2.6: Mapa mental dos trabalhos correlatos.....	21
Figura 3.1: Casos de uso do sistema de recuperação de informação em catálogos de metadados.....	22
Figura 3.2: Diagrama de sequência do caso de uso Gerar Base Indexada.	23
Figura 3.3: Diagrama de sequência do caso de uso consultar base indexada.....	23
Figura 3.4: Diagrama de sequência do caso de uso Avaliar Base Indexada.	24
Figura 3.5: Diagrama de componentes do sistema.	24
Figura 3.6: Diagrama de classes do módulo Geração da Base Indexada.	26
Figura 3.7: Diagrama de atividades do método GerarBaseIndexada.....	27
Figura 3.8: Diagrama de classes do módulo Avaliação da Base Indexada.....	27
Figura 3.9: Diagrama de atividades do método AvaliarBaseIndexada.	28
Figura 3.10: Diagrama de classes do módulo consulta da base indexada.	28
Figura 3.11: Diagrama de atividades do método ConsultarBaseIndexada.	29
Figura 3.12: Estrutura da arquitetura.....	30
Figura 4.1: Etapas de funcionamento do protótipo.....	31
Figura 4.2: Esquema conceitual do repositório.	32
Figura 4.3: Matriz de rastreabilidade: Classes X Bases.	33
Figura 4.4: Objeto da classe Recurso.	34
Figura 4.5: Trecho XML de um arquivo de metadado.	34
Figura 4.6: Objetos das classes Termo, Relacionamento e Atributo.	35
Figura 4.7: Trecho RDF/SKOS de um arquivo de tesouro.....	35
Figura 4.8: Tela inicial do protótipo.	36
Figura 4.9: Tela gerar base indexada do protótipo.	37
Figura 4.10: Tela avaliar base indexada do protótipo.....	38
Figura 4.11: Objetos das classes Termo, Relacionamento e Atributo.	39
Figura 4.12: Tela Consultar Base Indexada do protótipo.	40
Figura 5.1: IDE do projeto TERRANTAR.	42
Figura 5.2: Interface de pesquisa da IDE do projeto TERRANTAR.	45
Figura 5.3: Interface da ferramenta Tematres.....	45
Figura 5.4: Procedimento utilizado para evolução da base semântica.	46
Figura 5.5: Diagrama das atividades da etapa de publicação de dados e metadados.	47
Figura 5.6: Geração da Base Indexada.	48
Figura 5.7: Avaliação da Base Indexada.....	49
Figura 5.8: Visão parcial da árvore hiperbólica.....	50
Figura 5.9: Teste I: consulta com termos.....	51
Figura 5.10: Teste II: consulta com termos relacionados	52
Figura 5.11: Teste III: consulta com termos e especializações.....	53

LISTA DE TABELAS

Tabela 2.1: Uso de operadores e modificadores booleanos nas expressões de busca (N expressões = 51,473)	17
Tabela 5.1: Conteúdo da Base de Recurso	48

LISTA DE ABREVIATURAS

API - Application Programming Interface
DOM - Document Object Model
CONCAR - Comissão Nacional de Cartografia
FGDC - Federal Geographic Data Committee
ICA - International Cartographic Association
IDE - Infraestruturas de Dados Espaciais
ISO - International Standards Organization
OGC - Open Geospatial Consortium
RDF - Resource Description Framework
RI - Recuperação de Informação
SDI - Spatial Data Infrastructure,
SGBDR - Sistema de Gerenciamento de Banco de Dados Relacional
SIG - Sistemas de Informação Geográfica
SKOS - Simple Knowledge Organisation Systems
SRI - Sistemas de Recuperação de Informação
UML - Unified Modeling Language
W3C - World Wide Web Consortium
WCS - Web Coverage Service
WFS - Web Feature Service
WFS-T - Web Feature Service Transacional
WMS - Web Map Service
XML - eXtensible Markup Language
XSD - XML Schema Definition
XTM - XML *TOPIC MAPS*

RESUMO

SILVA, Odilon Corrêa da. M. Sc., Universidade Federal de Viçosa, dezembro de 2009. **Uso de tesauro e *topic maps* na recuperação de metadados em infraestruturas de dados espaciais: um estudo de caso do ecossistema antártico.** Orientador: Jugurta Lisboa Filho. Co-orientadores: José Luis Braga e Karla Albuquerque de Vasconcelos Borges.

A preocupação com fatores ambientais é uma realidade em todos os setores da sociedade, principalmente por parte de pesquisadores e profissionais que trabalham diretamente com o meio ambiente. Neste contexto, diversas pesquisas têm sido realizadas buscando respostas sobre os efeitos do aquecimento global no ecossistema antártico, gerando grande quantidade de dados geoespaciais. A disseminação e o compartilhamento destes dados têm enorme projeção se apoiados por um ambiente computacional. Este cenário apresenta muitos desafios de pesquisa em Ciência da Computação, tanto no nível físico (diversidade de estruturas de armazenamento) quanto conceitual (como a diversidade de perspectivas e de domínios de conhecimento). Infraestruturas de Dados Espaciais possibilitam a cooperação e o compartilhamento de dados geoespaciais; no entanto, o caráter aberto e distribuído dessas infraestruturas é um fator que aumenta a dificuldade de recuperação e compartilhamento desses dados. Pesquisas realizadas com o objetivo de recuperar informação e que utilizam artefatos semânticos como instrumentos de apoio foram analisadas. Entretanto, não foram encontrados trabalhos que abordam a recuperação de informação em catálogos de metadados combinado com mecanismos funcionais e intuitivos ao usuário. O objetivo desta dissertação é construir um mecanismo de busca que tem como foco principal a semântica, ou seja, o entendimento do contexto nas buscas a metadados é priorizado. O mecanismo proposto compreende um sistema de recuperação de informação em catálogos de metadados apoiado em artefatos de representação do conhecimento, como tesauros, *topic maps* e técnicas visuais. As simulações realizadas por meio de um estudo de caso permitiram verificar a viabilidade do sistema, no contexto de uma Infraestrutura de Dados Espaciais voltada para o ecossistema antártico. Também foi possível verificar a importância da capacitação dos usuários na produção de metadados geográficos. A análise e discussão dos resultados obtidos com as simulações mostraram que a eficiência do sistema de busca proposto está relacionada com a qualidade na produção dos metadados. Isso mostra a viabilidade de utilizar tesauros e *topic maps* como artefatos semânticos para a criação e recuperação de metadados.

ABSTRACT

SILVA, Odilon Corrêa da. M. Sc., Universidade Federal de Viçosa, december 2009. **The use of thesaurus and topic maps in the metadata recovery in spatial data infrastructures: a case study of the antartic ecosystem.** Advisor: Jugurta Lisboa Filho. Co-advisors: José Luis Braga and Karla Albuquerque de Vasconcelos Borges.

The preoccupation about environmental factors is a reality in all the sectors of society, especially when concerning researchers and professionals who work directly with the environment. In this context, several researches have been developed, trying to get answers about the effects of the global heating in the Antarctic ecosystem, which generates a big amount of geospatial data. The dissemination and sharing of this data can have enormous projection if supported by a computational environment. This setting presents many challenges in the Computer Science research, both in the physical (diversity of storage structures) and conceptual level (as the diversity of perspectives and knowledge domains). Infrastructures of Spatial Data enable the cooperation and geospatial data sharing; however, the open and distributed character of these infrastructures is a factor that increases the difficulty of retrieving and sharing these data. Researches developed with the aim of retrieving information, and which utilize semantic artifacts as supporting instruments, have been analyzed. However, there haven't been found any works that approached the information retrieval in catalogs of metadata combined with functional and intuitive mechanisms for the user. The aim of this dissertation is to build a search engine that has semantic as main point, which means, the understanding of the context in the metadata research is prioritized. The suggested mechanism comprises a system of data retrieval in metadata catalogs based on artifacts of knowledge representation, such as thesaurus, topic maps and visual techniques. The simulations performed through a case study enabled to verify the viability of the system, in the context of a Spatial Data Infrastructure regarding the Antarctic ecosystem. It was also possible to check the importance of the users' capacitating in the production of geographic metadata. The analysis and discussion of the results obtained with the simulations showed that the efficiency of the proposed search engine is linked to the quality of the metadata production. This shows the viability of utilizing thesauri and topic maps as semantic artifacts for the metadata creation and retrieval.

1 INTRODUÇÃO

A preocupação com fatores ambientais é uma realidade em todos os setores da sociedade. O planejamento sustentável de qualquer atividade, por exemplo planejamento urbano, reflorestamento e agricultura, requer, primeiramente, o conhecimento do meio ambiente em que esta atividade está inserida (BURROUGH e McDONNELL, 1998).

Para a caracterização ambiental de uma área, é necessário o conhecimento dos seus recursos naturais e o entendimento da interação e correlação entre eles (RESENDE et al., 1995). Os Sistemas de Informação Geográfica (SIG), juntamente com técnicas e produtos de sensoriamento remoto, podem auxiliar na busca desse conhecimento, bem como no gerenciamento e na atualização constante dos dados espaciais coletados (RESENDE et al., 1995; BURROUGH e McDONNELL, 1998).

A necessidade de compartilhar informações e a rápida evolução da Internet tornou-a um meio preferencial para a disseminação de dados espaciais. A alcançabilidade da Internet motivou o desenvolvimento de uma nova classe de sistemas de informação, com uma arquitetura diferenciada em relação a seus predecessores. Esse movimento se estende aos dados geográficos. Os principais fornecedores de SIG dispõem de alternativas para acesso a dados geográficos por meio da rede mundial (DAVIS e ALVES, 2005).

Em sua evolução, os SIG tornaram-se parte integrante da infraestrutura de sistemas de informação de muitas organizações. Como consequência ocorreu um aumento significativo no número e volume das fontes de dados espaciais disponíveis. Assim, o compartilhamento e a democratização dessa grande base de dados tornam-se cada vez mais importantes.

A disseminação e o compartilhamento de conjuntos de dados geográficos de origens diversas têm enorme projeção, se apoiados por um ambiente computacional em que os dados são livremente compartilhados de maneira integrada. Entretanto, segundo Davis e Alves (2005), essa não é a realidade atual, pois as organizações que têm a intenção de compartilhar dados normalmente se deparam com obstáculos significativos referentes aos formatos de codificação, armazenamento, parâmetros de qualidade, limitações de conteúdo, parâmetros de projeção cartográfica e, até mesmo, estruturas de dados.

Uma solução para minimizar esse problema está no uso de Infraestruturas de Dados Espaciais (IDE), que possibilitam a cooperação e o compartilhamento dos dados em questão (MAGUIRE e LONGLEY, 2005), pois o caráter aberto e distribuído das IDEs é um fator que aumenta a dificuldade de recuperação e compartilhamento de dados geoespaciais (ATHANASIS et al., 2008).

1.1 O problema e sua importância

A pesquisa sobre o ecossistema antártico tem grande importância, visto que essa é uma das regiões do planeta mais sensíveis às mudanças climáticas globais. A Antártica é o termômetro da saúde do planeta e onde primeiro se percebem as alterações que ele vem sofrendo (MMA, 2007). O continente detém 90% do gelo e 80% da água doce da Terra, além de riquezas minerais incalculáveis. O vasto manto de gelo antártico é o principal sorvedouro do calor terrestre, por isso tem um papel essencial no sistema climático global.

A busca por respostas sobre os efeitos do aquecimento global no ecossistema antártico requer correlacionar, por exemplo, pesquisas sobre solos, flora, qualidade da água, comportamento das correntes marítimas e do meio ambiente antártico. Neste contexto, o conhecimento gerado necessita estar disponível aos pesquisadores, a fim de permiti-los analisar as características desses elementos e suas interações.

Em geral, a pesquisa em IDE baseia-se principalmente em buscas que usam palavras-chave, coordenadas espaciais, classificação temática ou temporal. Com o resultado da pesquisa, o usuário pode acessar a fonte dos dados para visualizá-los ou obtê-los, analisar seus metadados ou iniciar uma nova pesquisa. Essa abordagem apresenta algumas dificuldades, especialmente para os usuários inexperientes, que podem não saber quais palavras-chave usar, como preencher adequadamente o formulário de consulta ou definir quantos critérios serão utilizados (HOCHMAIR, 2005). Essas limitações no processo de recuperação do conhecimento podem prejudicar buscas mais complexas.

Para contornar esses problemas, buscou-se uma nova abordagem na recuperação e no compartilhamento de informações geoespaciais, que, juntamente com as questões decorrentes da interação entre componentes de uma IDE, torna-se tema central da Web Semântica Geoespacial (EGENHOFER, 2002). Nesse contexto, surge a seguinte questão: como produzir um mecanismo que seja influenciado por artefatos semânticos na recuperação de dados em catálogos de metadados de uma IDE?

Tornando isso possível, seriam alcançados ganhos na forma como os resultados gerados por uma comunidade possam estar disponíveis aos usuários e pesquisadores, a fim de permiti-los analisar as características desses elementos e suas interações.

Por fim, mas não menos importante, algumas ações foram propostas no seminário “Grandes Desafios da Computação no Brasil: 2006-2016”, organizado pela Sociedade Brasileira de Computação. Uma questão levantada é a necessidade de integração dos pesquisadores da área de Computação em projetos multidisciplinares. Uma proposta com cinco grandes desafios para as pesquisas em Computação para a próxima década foi elaborada:

1. Gestão da Informação em grandes volumes de dados multimídia distribuídos;
2. Modelagem computacional de sistemas complexos artificiais, naturais e socioculturais e da interação homem-natureza;
3. Impactos para a área da computação da transição do silício para novas tecnologias;
4. Acesso participativo e universal do cidadão brasileiro ao conhecimento; e
5. Desenvolvimento tecnológico de qualidade: sistemas disponíveis, corretos, seguros, escaláveis, persistentes e ubíquos.

Os problemas apresentados nesta seção estão inseridos no contexto dos desafios 1 e 4, o que corrobora a importância deste trabalho.

1.2 Objetivos

O objetivo geral desta pesquisa é minimizar alguns problemas relacionados à semântica na recuperação de informação em catálogos de metadados em uma Infraestrutura de Dados Espaciais utilizando-se de Tesauro e *Topic maps*. Para isso, a pesquisa tem como objetivos específicos:

1. Investigar IDE existentes e seus mecanismos de recuperação de informações em catálogos de metadados;
2. Investigar uma estrutura para representar e armazenar a base semântica do conhecimento em estudo;
3. Investigar e especificar uma forma de integrar a base semântica aos metadados do catálogo de uma IDE;
4. Definir um procedimento para auxiliar o usuário a gerenciar a catalogação de metadados e a evolução do conteúdo da base semântica;
5. Especificar e desenvolver um sistema de recuperação de informação em

- catálogos de metadados que seja influenciado pela integração entre a base semântica e metadados do catálogo da IDE; e
6. Verificar a viabilidade do sistema proposto por meio de um estudo de caso, no contexto de uma IDE.

1.3 Organização da dissertação

Esta dissertação está organizada da seguinte forma:

- Capítulo 1. Apresenta o problema, sua importância e os objetivos do trabalho.
- Capítulo 2. Descreve o referencial teórico, resultado da pesquisa bibliográfica sobre as áreas e tecnologias relacionadas ao trabalho.
- Capítulo 3. Apresenta uma proposta de arquitetura para um sistema de recuperação de informação em catálogo de metadados de uma IDE.
- Capítulo 4. Detalha os aspectos de implementação relacionados ao protótipo do sistema proposto no Capítulo 3.
- Capítulo 5. Descreve um estudo de caso realizado com dados reais relacionados ao ecossistema antártico.
- Capítulo 6. Apresenta as conclusões e possíveis desdobramentos desta dissertação.
- Ao final do documento, é apresentado um apêndice contendo um glossário com a definição de todos os elementos do tesauro utilizado no Capítulo 5.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta uma visão geral sobre IDE e, em particular, sobre os aspectos da utilização de metadados espaciais (Seção 2.1). Em seguida, são descritos conceitos utilizados ao longo desta dissertação. Esta pesquisa fundamenta-se em duas áreas, representação do conhecimento por tesauros, ontologias ou *topic maps* (Seção 2.2) e recuperação de informação (Seção 2.3). Os trabalhos relacionados, apresentados na Seção 2.4, descrevem outras pesquisas que influenciaram e contribuíram para o desenvolvimento desta dissertação.

2.1 Infraestrutura de dados espaciais

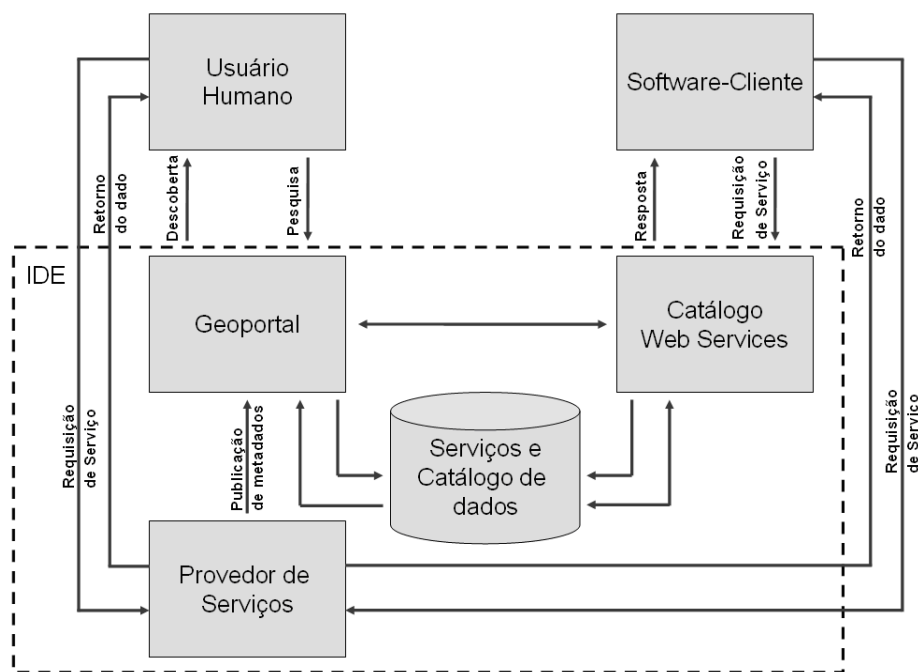
A partir de algumas definições de padrões de intercâmbio, muitas agências nacionais de mapeamento criaram *Clearinghouses* de dados espaciais. Trata-se de um aparato computacional baseado na Internet que facilita o acesso a dados espaciais, pela criação de um local centralizado onde podem ser encontrados dados de diversas fontes, e onde estão disponíveis serviços complementares, como busca, visualização, transferência e manipulação de dados espaciais (CROMPVOETS et al., 2004).

Segundo Crompvoets et al. (2004), há diferentes entendimentos quanto ao conceito de *Clearinghouse*. Mais recentemente, *Clearinghouses* têm sido descritas de forma muito similar a um portal Web, ou seja, um local onde diversos serviços de uso comum são oferecidos, ou um concentrador por meio do qual esses serviços são acessados. Um estudo comparando características de *Clearinghouses* relata uma crescente insatisfação entre seus usuários com relação às suas respectivas funcionalidades. O estudo indica, ainda, que o foco deveria ser o usuário, em vez de seus dados (CROMPVOETS et al., 2004).

Buscando um novo foco no desenvolvimento de dados para usuários e aplicações, muitas *Clearinghouses* evoluíram para o que Masser (1999) chama de “primeira geração de IDE”. De acordo com Maguire e Longley (2005), a expressão “Infraestrutura de Dados Espaciais – IDE” (do inglês Spatial Data Infrastructure, SDI) foi proposta em 1993 pelo *Mapping Sciences Committee*, do *U.S. National Research Council*, sendo usada inicialmente para descrever o provimento de acesso padronizado à informação geográfica.

A primeira geração de IDE volta sua atenção para a garantia de um escopo

temático amplo, o que está de acordo com objetivos que permitem uma analogia entre IDE e outros tipos de infraestruturas: fomentar o desenvolvimento econômico pela garantia de acesso a serviços e produtos disponíveis publicamente e de interesse para um grupo significativo de pessoas. Davis e Alves (2005) propõem uma arquitetura para o desenvolvimento de IDE na qual os dados são providos por diferentes serviços de informação, formando assim uma infraestrutura orientada a serviços (Figura 2.1), o que pode ser chamado de “segunda geração de IDE”. A arquitetura proposta por Davis e Alves (2005) será utilizada como base no desenvolvimento da IDE do estudo de caso desta dissertação.



Fonte: Davis e Alves (2005)

Figura 2.1: Estrutura de IDE orientada a serviços.

Para Davis e Alves (2005), o foco no conceito de IDE é uma consequência natural da evolução da Web e de sua arquitetura, diante das diretrizes que vêm sendo traçadas por seu órgão normatizador, o *World Wide Web Consortium* (W3C). A principal mudança na Web é que a apresentação deixou de ser o foco central para dividir seu lugar com o conteúdo, fator que permite a estruturação semântica dos dados. Essa mudança de foco foi cristalizada no conceito da Web Semântica, “uma extensão da Web na qual a informação possui um significado bem definido, possibilitando que computadores e pessoas trabalhem em cooperação” (BERNERS-LEE, 2001).

Segundo Nebert et al. (2004), o termo IDE é usado para designar um conjunto de

políticas, tecnologias, padrões e recursos humanos necessários (principalmente as organizações produtoras de cartografia) para que se promovam a disponibilidade e o acesso a dados espaciais. O trabalho descreve um conjunto de princípios, normas e protocolos, para que IDE adjacentes troquem informações. Os autores apresentam os aspectos necessários para se iniciar a construção de uma IDE, os quais incluem dados espaciais, metadados, catálogos, visualização de dados online, acesso aos dados, geoserviços, capacitação e políticas públicas.

Em ambientes onde há grande heterogeneidade nos dados tratados, o uso de metadados permite relacionar dados por meio de uma classificação que seja comum entre os diversos tipos de dados disponíveis, a fim de melhorar a integração e o compartilhamento destes dados. De acordo com Casanova et al. (2005), os metadados tratam a interoperabilidade em nível de gerenciamento da informação, facilitando a recuperação e o uso de um dado ou informação contida em um ambiente computacional.

Atualmente, existem diversos padrões de metadados espaciais que são utilizados para documentar coleções de dados geoespaciais. Dentre estes, as experiências do governo americano com o *Content Standards for Digital Geospatial Metadata* (CGSDM) do *Federal Geographic Data Committee* - FGDC (FGDC, 1998) e a norma ISO/TC 211:19115 (ISO 19115, 2003) são amplamente utilizados. A norma ISO/TC 211:19115 foi desenvolvida pela *International Standards Organization* (ISO), órgão responsável pela definição de um conjunto estruturado de padrões de informações associadas direta ou indiretamente à localização espacial.

A norma ISO/TC 211:19115 serviu de base para a especificação do Perfil de Metadados Geoespaciais do Brasil (Perfil MGB)¹, em uma iniciativa para ordenar a geração, o armazenamento, o acesso, o compartilhamento, a divulgação e o uso dos dados geoespaciais no território nacional. Essa especificação surgiu de um trabalho elaborado por um dos comitês especializados da Comissão Nacional de Cartografia (CONCAR), o Comitê de Estruturação de Metadados Geoespaciais (CEMG), criado especificamente para esse fim.

Esses padrões normalmente estão baseados nos elementos de qualidade definidos pela *International Cartographic Association* (ICA) (GUPTILL e MORRISSON, 1997) e disponibilizam informações para o conhecimento dos dados existentes, o enquadramento dos dados em determinadas aplicações e as condições de acesso e

¹ Perfil MGB - http://www.concar.ibge.gov.br/arquivo/Perfil_MGB_Final_v1_homologado.pdf

transferência para o usuário. A utilização de um modelo de referência de metadados para os dados geográficos possibilita a descrição de suas propriedades segundo diversas abordagens descritivas e este modelo de referência deve ser flexível o bastante para se ajustar facilmente aos novos requisitos das aplicações dos usuários. A norma ISO/TC 211:19115 (ISO 19115) é utilizada como padrão para catalogar os metadados da IDE do estudo de caso desta dissertação. Os metadados do estudo de caso são armazenados em XML² (*eXtensible Markup Language*), de acordo com o XSD (*XML Schema Definition*) da norma ISO/TC 211:19115 (ISO 19115).

Em Nogueras-Iso et al. (2005), são levantadas algumas questões relacionadas ao desenvolvimento de IDE, dando ênfase ao papel desempenhado por metadados. Os autores defendem a importância dos metadados e abordam questões relacionadas com a interoperabilidade entre padrões: ISO, FGDC e Dublin Core, além de descrever benefícios da utilização de tesauros e ontologias na recuperação de informação em catálogos de metadados. Outro problema destacado é a pouca ou total falta do relacionamento entre Bibliotecas Digitais e IDEs. Os autores alertam para o fato de que experiências acumuladas nessa área podem fortalecer o desenvolvimento de conceitos em muitos aspectos de uma IDE. Nesse sentido, eles sugerem que Bibliotecas Digitais poderiam ser consideradas como uma parte de IDE.

2.2 Representação do conhecimento

A compreensão semântica dificulta a interação dos usuários humanos de uma IDE, já que as visões de mundo podem influenciar, por exemplo, na consulta a metadados. A compreensão da semântica deve ser um requisito para uma IDE que pretende dar suporte à pesquisa sobre o meio ambiente (CRAGLIA e GOODCHILD, 2008; DAVIS JR. et al., 2009; DAVIS JR. e FONSECA, 2009). Com isso, as várias visões envolvidas nas questões ambientais necessitam de mecanismos que facilitem a navegação entre os conjuntos de conceitos, ou termos, que caracterizam cada grupo de usuário (DAVIS JR. et al., 2009).

De acordo com Davenport e Prusak (1998), para disponibilizar o conhecimento em um formato que seja acessível ao usuário, é necessário definir uma codificação. Um aspecto fundamental para atingir eficientemente esses resultados é a forma como esses conhecimentos são representados.

Para Garshol (2004), uma das preocupações de um arquiteto de informação é

² XML: Schema Part 0: Primer Second Edition - <http://www.w3.org/TR/xmlschema-0/>

criar sites onde os usuários realmente possam encontrar o que estão procurando. Para criação e organização do site, o arquiteto de informação utiliza-se de diversas técnicas e tecnologias, mas seus principais instrumentos estão relacionados à organização da informação, estes têm origem em outras disciplinas. Algumas dessas técnicas vêm da biblioteconomia, como vocabulários controlados, taxonomias e tesauros.

A seguir, são discutidos em especial tesauros, ontologias e *topic maps*, explorando suas contribuições específicas para a representação e recuperação de informação.

2.2.1 Tesouro

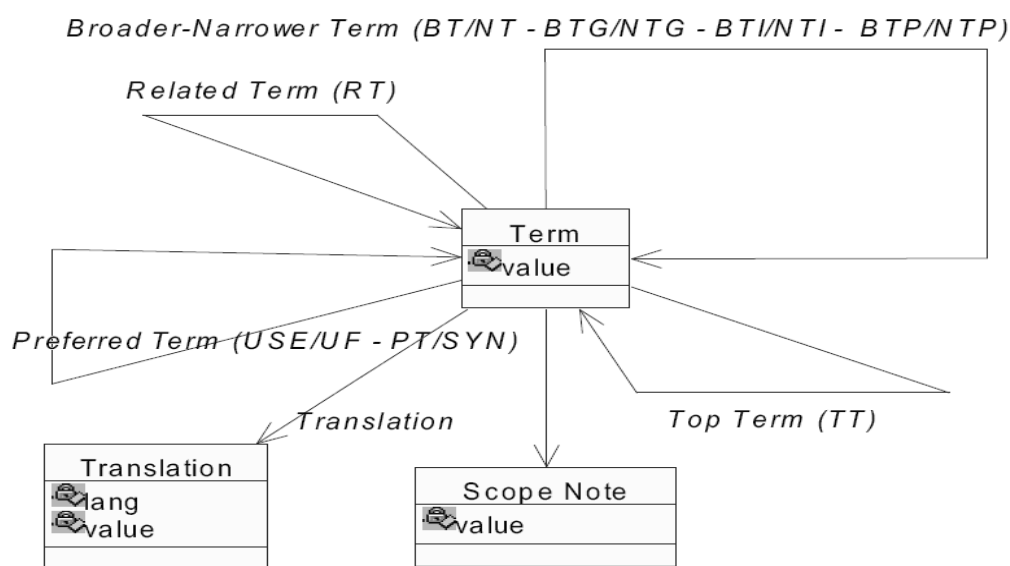
O termo tesouro, segundo Dodebei (2002), origina-se do grego *thesaurós* e significa tesouro ou repertório. Este termo se popularizou com a publicação do dicionário analógico de Peter Mark Roget (1852) apud Dodebei (2002), intitulado “*Thesaurus of english words and phrases*”. Roget chamou de tesauros seu dicionário de palavras, uma vez que o termo também designa vocabulário, dicionário ou léxico. O dicionário do autor diferenciava-se dos outros por ser um vocabulário organizado de acordo com o seu significado e não pela ordem alfabética e teve o mérito de estabelecer a denominação para vocabulários que relacionam seus termos por meio de algum tipo de relação de significado.

Segundo Foskett (1972) apud Dodebei (2002), o principal objetivo do tesouro é o controle terminológico e, em concordância com Lancaster (1972) apud Dodebei (2002). Foskett lista como suas principais funções o controle de sinônimos e quase sinônimos; a distinção de homógrafos; a promoção de facilitadores para condução da busca por meio dos termos relacionados e das referências cruzadas, melhorando a consistência da indexação e transportando a linguagem de busca para a de indexação; e a redução do tempo e o aumento da eficiência nas tarefas de indexação e recuperação de informações.

O tesouro foi adotado na área de documentação, associado à forma de organização do vocabulário de indexação/recuperação, utilizado amplamente na área da Ciência da Informação (DODEBEI, 2002). De fato, o emprego de tesauros nas tarefas de indexação e recuperação de informações busca resolver o problema da alocação de documentos em classes de assuntos, não só por sua capacidade de controlar o vocabulário, mas porque é um instrumento que relaciona os descritores/termos de forma mais consistente, apresentando uma estrutura sintética simplificada e uma complexa rede de referências cruzadas. Isso permite ao especialista localizar, com mais facilidade, a palavra-chave

requerida para uma busca.

O W3C propõe um padrão baseado em RDF (*Resource Description Framework*) chamado SKOS³ (*Simple Knowledge Organisation Systems*) para representação de tesouros e outros tipos similares de sistema de organização de conhecimento (SKOS, 2009). As especificações para construção de tesouro estão definidas pelas normas ISO 2788 (ISO, 1986) e ISO 5964 (ISO, 1985). A Figura 2.2 ilustra um modelo para representação das componentes de um tesouro.



Fonte: Miguel (2009)

Figura 2.2: Modelo para representação de um tesouro.

Pode ser observado na Figura 2.2 que esse modelo de representação tem suporte *multilíngue*, atendendo às especificações da ISO 5964 (ISO, 1985). Os símbolos do modelo são apresentados a seguir:

- NA – Nota de Alcance (*Scope Note*): define, explica ou limita o significado do descritor para finalidades de indexação;
- UP – Usado Para (*User For*): indica os sinônimos ou não-descritores; eles não são válidos para a indexação;
- TG – Termo Genérico (*Broader Term*): assinala o termo mais amplo ao qual pertence o descritor; e
- TE – Termo Específico (*Narrower Term*): indica os termos específicos (tipos ou classes) associados ao descritor;

³ <http://www.w3c.org/2004/02/skos/>

- TR – Termo Relacionado (*Related Term*): indica os termos associados semanticamente com o descritor.

Existem vários tesouros criados e mantidos por organizações não governamentais disponíveis para consulta e acesso ao seu conteúdo na Internet, alguns exemplos citados e utilizados nesta dissertação são:

- AGROVOC⁴: é um tesouro de termos agrícolas *multilíngue* mantido pela Organização para a Agricultura e a Alimentação (FAO, sigla de *Food and Agriculture Organization*), que é uma organização das Nações Unidas cujo objetivo declarado é elevar os níveis de nutrição e de desenvolvimento rural; e
- **GEneral Multilingual Environmental Thesaurus** (GEMET)⁵: é um tesouro de termos para o meio ambiente, desenvolvido pelo Centro Temático Europeu (*European Topic Centre*) e mantido atualmente pela Rede Europeia de Observação e Informação do Meio Ambiente (EIONET, sigla de *European Environment Information and Observation Network*).

Os tesouros AGROVOC e GEMTE estão disponíveis ao público para consulta por meio de um sítio e serviço Web. As bases de dados destes tesouros também estão disponíveis em vários formatos para *download*.

Dodebei (2002) considera que a determinação do universo conceitual pode ser obtida tanto por conhecimento prévio do domínio quanto por observação do campo conceitual, reconhecendo, no campo da metodologia, a indução e a dedução como os processos de escolha para a composição desse universo. No processo indutivo, chamado de “Empírico” por Lancaster (1972) apud Dodebei (2002) e “Analítico” pelo *American National Standard Institute* (ANSI), a obtenção da terminologia faz-se mediante a identificação de termos coletados na literatura. Já no processo conceitual, a obtenção da terminologia se faz por consenso de peritos no assunto, o método é chamado também de *Gestalt*, em contraposição a “Analítico”, ou ainda *Committee Approach*, em contraposição a “Empírico” (Dodebei, 2002). Segundo Dodebei (2002), os dois procedimentos derivam, por sua vez, de dois princípios que regem o levantamento dos termos e que são, respectivamente, a Garantia Literária (*Literary Warrant* ou *Bibliographic Warrant*), usada por Hulme (1950) apud Dodebei (2002) e o Endosso do Usuário (*User Warrant* ou *Personal Warrant*), usado por Lancaster (1972) apud

⁴ http://www.fao.org/aims/ag_intro.htm

⁵ <http://www.eionet.europa.eu/gemet>

Dodebei (2002). Os métodos citados são utilizados na fundamentação teórica, criação e evolução do tesauro utilizado nesta dissertação.

Existem vários sistemas de gestão de linguagens documentárias, ferramentas que auxiliam na criação e manutenção de tesouros. Em Campos et al. (2006), os autores fazem um estudo comparativo de *softwares* de construção de tesouros. Após analisar os critérios de avaliação utilizados pelos autores, em conjunto com as necessidades de criação e evolução do tesauro desta dissertação, optou-se pelo software Tematres⁶.

O Tematres é uma aplicação Web para a gestão de linguagens documentárias com suporte *multilíngue* e tem interoperabilidade com arquivos no formato SKOS. Outro fator determinante é o fato de ser uma aplicação Web, o que facilita a cooperação entre possíveis colaboradores na criação e evolução deste tesauro.

2.2.2 Ontologia

A definição de ontologia proposta por Gruber (1995) é frequentemente encontrada na literatura: “Ontologia é uma especialização formal e explícita de uma conceitualização compartilhada”. A consideração mais importante nessa afirmação é a noção de conceitualização. A conceitualização pode ser desenvolvida contendo terminologias e vocabulários, estabelecendo propriedades e permitindo que o conhecimento seja reutilizado, evitando o retrabalho ou a redescoberta de terminologias equivalentes (GUARINO, 1998).

Com uma longa história em filosofia, o termo ontologia foi incorporado ao vocabulário da Inteligência Artificial, Engenharia de Conhecimento e Representação do Conhecimento, entre outras áreas da Ciência da Computação. Nas últimas duas décadas, as pesquisas de Inteligência Artificial vêm se servindo de ontologias no processo de descrição formal das coisas do mundo, algo fundamental para que sistemas inteligentes possam atuar e ponderar sobre o mundo em que se propõem operar (WELTY e GUARINO, 2001).

Segundo Sowa (2001), pesquisadores da área de representação do conhecimento denominam ontologia terminológica como estrutura capaz de capturar e registrar a semântica dos elementos terminológicos de determinado contexto. O autor difere uma ontologia terminológica e formal pelo grau, podendo uma ontologia terminológica evoluir para uma ontologia formal à medida que mais axiomas são acrescentados. A pesquisa do Sowa tem relação e concordância com a de outros pesquisadores (JASPER

⁶ sourceforge.net/projects/tematres/

e MIKE USCHOLD, 1999; FENSEL, 2001), que consideram os tesouros como ontologias simples (terminológica), uma vez que uma ontologia complexa (formal) exige uma riqueza maior de relações (axiomas) do que as apresentadas em um tesouro.

Em Moreira et al. (2004), é colocada a seguinte questão: “Seriam tesouros e ontologias termos que denotam o mesmo instrumento ou denotariam objetos diferentes e, portanto, seu uso indistinto seria inadequado?”. Esse trabalho está relacionado com outra pesquisa (MOREIRA, 2003) e faz uma revisão e discussão da literatura sobre tesouros e ontologias, com o objetivo de contribuir com o esclarecimento e entendimento desses dois conceitos. Um fato observado pelos autores é que muitos instrumentos de suporte a sistemas de recuperação e troca de informação ou conhecimento têm sido denominados como “ontologias”, o que seria um equívoco causado pela falta de entendimento sobre o que caracteriza cada um desses instrumentos.

Ontologias são modelos de representação do conhecimento que, a exemplo dos tesouros, são utilizados para representar e recuperar conhecimento/informação por meio de uma estrutura conceitual. Não é objetivo desta dissertação entrar em questões polêmicas relacionadas às ontologias, mas a revisão bibliográfica realizada constatou que ontologias vão além dos instrumentos mencionados até então. Historicamente, as ontologias (empregadas e adotadas pela Filosofia) apresentam idade muito inferior em relação aos tesouros. Na ciência da computação, o termo vem sendo mais aceito e desenvolvido cientificamente, pois se trata de uma ferramenta cujo objetivo é representar e recuperar informação dentro do ambiente digital. É fato que ainda se busca um consenso com relação às definições e aplicações de ontologias.

Segundo Maedche e Staab (2001), o processo de construção de ontologias é complexo e longo, envolvendo o levantamento de conceitos e termos de interesse em um determinado domínio, de acordo com o propósito que se deseja atingir. Em resposta a essa dificuldade, a comunidade científica vem explorando novas técnicas e abordagens que possam reduzir esse esforço, por exemplo, criar uma ontologia a partir de um tesouro existente.

Portanto, visto que o tesouro é um componente de extrema relevância para um sistema de recuperação de informação, além do consenso que há em sua aplicação em relação às ontologias. Esses fatos, em conjunto com seus aspectos de implementação, apoiaram a decisão de adotar o tesouro como instrumento de representação da base semântica do sistema de recuperação de informação desta dissertação. Um futuro

trabalho que tenha o objetivo de criar uma ontologia para o ecossistema antártico (âmbito do estudo de caso desta dissertação, ver Capítulo 5) pode se beneficiar do tesouro criado por uma comunidade que trabalhe com esse contexto.

2.2.3 *Topic maps*

No ambiente digital que vem se configurando nas últimas décadas, os acervos de objetos digitais multiplicam-se tanto no que se refere à sua tipologia quanto à sua complexidade. Nesse cenário, textos, imagens, sons, vídeos, páginas Web e diversos outros objetos digitais requerem diferentes tipos de tratamento e representação para uma recuperação de informação eficaz (BURKE, 1999).

Segundo Sigel (2000), a gestão de conhecimento busca organizar e otimizar repositórios de conhecimento para que o usuário consiga recuperar informações de forma eficaz. De acordo com Rath (2003), “falar sobre estrutura de conhecimento é falar sobre *topic maps*”. Conforme a ISO 13250 (ISO, 2002), *topic maps* é uma norma internacional unificada para descrever estruturas de conhecimento e formalizar a sua associação com recursos de informações. *Topic maps* podem ser vistos como um paradigma que permite organizar e recuperar informações armazenadas em IDE. Sua adoção permite, por exemplo, a criação de um sistema de recuperação de informação em catálogos de metadados.

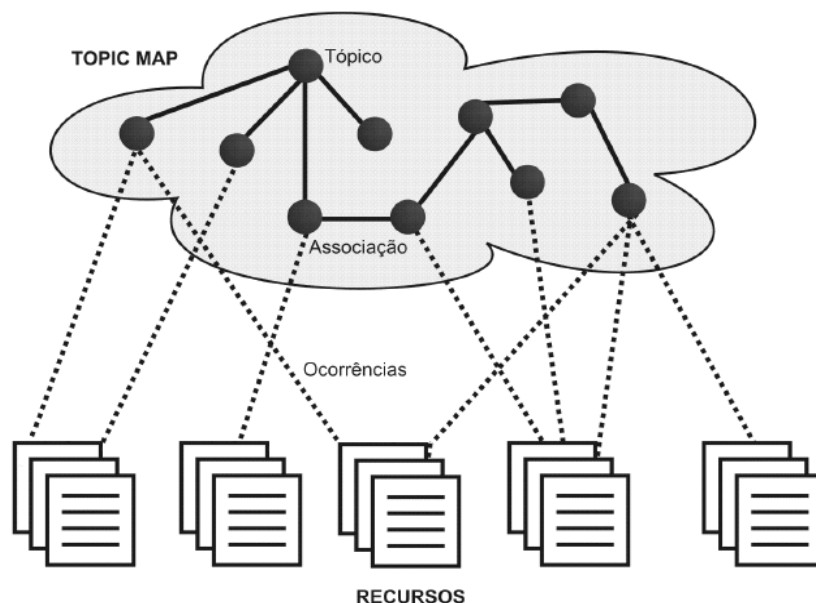
Topic maps são similares, em muitos aspectos, às redes semânticas, mapas conceituais e mapas mentais, embora somente os *topic maps* estejam padronizados, fator que permite futuras implementações de interoperabilidade. Sua estrutura está formalizada como norma ISO 13250 e especificação XTM (*XML Topic Maps*) (ISO, 2002). As formas de representação do conhecimento por meio de tesouros e de ontologias, que foram apresentadas nessa seção, influenciaram a criação da norma *Topic maps* (ISO, 2002). Sendo assim, pode-se dizer que um *topic maps* é um mapa que representa o conhecimento encontrado em um domínio, além de poder apresentar os conceitos relevantes e os relacionamentos entre eles, de forma similar ao que é feito no tesouro. Outra característica importante de um *topic maps* são as ocorrências entre os conceitos e recursos de informação.

Topic maps são compostos por três conceitos (Figura 2.3): Tópico, Associação e Ocorrência, representados pela sigla TAO (PEPPER, 2000). Apesar da simplicidade, esses elementos têm alto poder descritivo. Eles são descritos a seguir:

- **Tópico:** pode representar um tema em um domínio de aplicação, como uma

pessoa, uma entidade, um conceito etc. A rigor, o termo "tópico" refere-se a um elemento do *topic maps* que representa um tema. Os tópicos apresentam relações do tipo classe/subclasse. Por exemplo, “Instituição” é um supertipo de “Universidade”, que por sua vez é um supertipo de “Universidade Federal de Viçosa”;

- **Associação:** uma associação permite descrever relacionamentos entre tópicos. Uma associação é um elemento de vínculo que define um relacionamento entre dois ou mais tópicos. Exemplo de associação: os tópicos “Instituição” e “Evento” estão vinculados pela associação “Organizar”, que é representada pelos papéis “Organiza” e “Organizado-por”;
- **Ocorrências:** um tópico pode estar ligado a um ou mais recursos de informação, os quais são relevantes para o tópico em questão de algum modo. Os recursos são chamados ocorrências de tópicos. Uma ocorrência é uma conexão que relaciona um tópico com um recurso. Para exemplificar, considerem-se os recursos sítio (<http://www.ufv.br>) e ano de fundação (1922) como ocorrências do tópico "Universidade Federal de Viçosa".



Fonte: Garshol (2002)

Figura 2.3: Elementos da norma *topic maps*.

Ahmed (2003) apresenta um padrão de projeto para *topic maps* baseado em ontologias. Este método pode ser usado para representar tesouros em *topic maps* seguindo alguns princípios, propostos por Garshol (2004). O padrão de projeto proposto por Ahmed (2003) e adaptado por Garshol (2004) pode ser aplicado, por exemplo, na

criação de um sistema que permita processar e indexar os termos de um tesauro e palavras-chave de metadados geoespaciais, gerando o *topic maps*. Nesse contexto, a criação de uma base semântica e implementação de uma interface para recuperar informação em catálogos de metadados usando *topic maps* pode melhorar as pesquisas e contribuir para a recuperação de metadados em IDE.

2.3 Recuperação da informação

Segundo Kuhlthau (1990), uma busca de informação começa com o problema do usuário. A lacuna entre o conhecimento que o usuário detém sobre o problema ou tópico e o que usuário precisa saber para resolver o problema constitui a necessidade de informação. O termo “Recuperação de Informação” atribuído a sistemas computacionais é bastante questionado, muitos autores preferem o termo “Recuperação de Documento” (“*Document Retrieval*”) ou “Recuperação de Texto” (“*Text Retrieval*”) (Ferneda, 2003). Nota-se que os sistemas não recuperam “informação”, mas sim documentos ou referências cujo conteúdo pode ser relevante para a necessidade do usuário. Assim, será utilizada a designação “Recuperação de Informação” (RI) neste trabalho, ficando subentendido que se trata de uma “informação” contida nos documentos ou textos recuperados.

Os princípios básicos do modelo probabilístico para Sistemas de Recuperação de Informação (SRI) foram lançados por Maron e Kuhns (1960) apud Ferneda (2002) e definidos formalmente por Robertson e Jones (1976) apud Ferneda (2002). Os primeiros SRI baseavam-se na frequência das palavras no texto. Embora a utilização de métodos mais sofisticados tenha surgido em consequência de uma evolução natural dos modelos matemáticos na busca de um aprofundamento semântico no tratamento textual, pesquisas utilizando modelos estatísticos continuaram gerando novos modelos e aperfeiçoando antigas ideias. É o caso do modelo booleano e de diversos outros modelos que foram atualizados tendo em vista as necessidades e urgência de métodos de recuperação para a Web (FERNEDA, 2003).

Como notado por Ferneda (2003), a RI, na maior parte das bases de dados, dá-se através dos mecanismos que empregam os princípios da álgebra booleana e da teoria de conjuntos, o que se aplica também para os mecanismos de busca disponíveis na Internet. A presença dos mecanismos booleanos nesses sistemas não tem feito com que estes fiquem necessariamente mais fáceis de serem usados pelo usuário final. Este fato é demonstrado nas pesquisas de Savoy e Picard (2001) e Jansen et al. (2000), que

apontam a baixa utilização dos operadores booleanos pelos usuários das máquinas de busca da Web. Entre os operadores usados, o *AND* (E) é o mais utilizado, como apresentado na tabela 2.1.

Tabela 2.1: Uso de operadores e modificadores booleanos nas expressões de busca (N expressões = 51,473)

Operador ou modificador	Número de expressões	Porcentagem das expressões
AND	4.094	8,00
OR	177	0,34
AND NOT	105	0,20
()	273	0,53
“ ”	3.282	6,00

Fonte: Jansen et al. (2000, p.217)

O baixo uso dos recursos da álgebra booleana para a RI deve-se a vários fatores. A primeira limitação é que a recuperação booleana cria dois subconjuntos discretos, os que satisfazem a expressão de busca e aqueles que não a satisfazem. De modo semelhante, na recuperação booleana, não há mecanismos pelos quais o usuário possa atribuir a importância relativa dos diferentes termos da expressão, dado o fato de que todos os termos da busca booleana têm peso um ou zero na expressão. Outra limitação se relaciona à natureza ambígua da linguagem natural em que a necessidade de informação é expressa, implicando a possibilidade de recuperação de muitos registros irrelevantes (ALLEN, 2001; INGWERSEN e WILLETT, 1995).

O processo de recuperação de informação consiste em identificar, no conjunto de documentos de um sistema, quais deles atendem à necessidade de informação do usuário (BURKE, 1999). Segundo Grand e Soto (2002), técnicas visuais de pesquisa facilitam a percepção do ambiente pelo usuário. Para os autores, o recurso de navegação é essencial, pois ajuda o usuário a construir sua própria representação do ambiente.

A árvore hiperbólica é um exemplo de tecnologia para visualização e navegação baseada na técnica foco+contexto (LAMPING, 1995). Essa técnica pode ser utilizada para representação gráfica de *topic maps*, o que proporciona ao usuário maior facilidade de visualização e navegação pelo conteúdo informacional, sem que ele precise conhecer exatamente o termo desejado. Ao utilizar árvore hiperbólica para representar *topic maps*, o tópico posicionado pelo usuário no foco aparece maior. Na navegação, o aparecimento de tópicos periféricos cresce exponencialmente à medida que o *topic*

maps é deslocado. Devido a estas características, a navegação hiperbólica, em conjunto com tesauros, *topic maps* e métodos de RI, apresenta uma boa alternativa para a visualização e recuperação de grandes conjuntos de dados, por exemplo catálogos de metadados de uma IDE.

2.4 Trabalhos correlatos

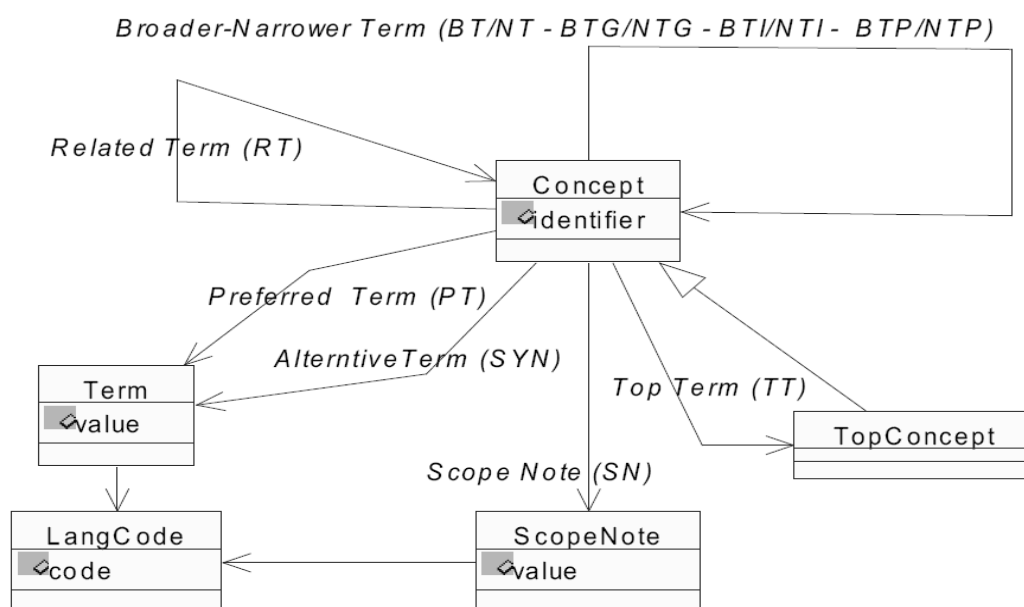
Nesta seção, são apresentados alguns trabalhos que influenciaram no amadurecimento da idéia de um SRI, que visa auxiliar no processo de RI em catálogos de metadados de uma IDE. A seguir, cada um desses trabalhos é discutido sob o ponto de vista da sua contribuição para o desenvolvimento desta dissertação.

O SBICAFÉ (PONTES et al., 2001) é uma ferramenta de gestão do conhecimento que foi desenvolvida na forma de um portal na Internet e conta com um tesouro *multilíngue* de terminologia agrícola. O projeto utiliza-se do tesouro AGROVOC para auxiliar os usuários no processo de consulta à sua base de dados. O SBICAFÉ foi desenvolvido para disseminar os resultados de pesquisas no âmbito do Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café e das instituições participantes. Os autores citam algumas funcionalidades a serem desenvolvidas no sistema. Essas funcionalidades estão relacionadas ao uso de ontologias para melhorar os recursos de busca e navegação no portal. Os principais objetivos do uso de ontologias são: fornecer mecanismos de navegação funcionais e intuitivos, formar corretamente a expectativa dos usuários em relação ao conteúdo disponível e, principalmente, fornecer mecanismos de busca amigáveis, interativos e de fácil uso para leigos.

No contexto das IDEs, os pesquisadores Aditya e Kraak discutem e propõem o desenvolvimento de interfaces e mecanismos de RI em catálogos de metadados (ADITYA e KRAAK, 2007; ADITYA, 2007). Para os autores, é essencial a evolução das interfaces de pesquisa das IDE, por isso eles propõem a utilização de mapas e gráficos para apoiar a descoberta de recursos geoespaciais. Nesta proposta, os metadados geoespaciais são indexados com base em sua ligação com um mapa temático, estes que são agrupados pela taxonomia *TopicCategory* da ISO19115 (ISO/TC211 2003). Os produtores de dados geoespaciais ficam responsáveis por especificar qual mapa temático é mais adequado para indexar seus dados, essa tarefa é obrigatória e está vinculada ao registro do metadado. Com os mapas temáticos indexados, é possível disponibilizar no portal da IDE uma interface de pesquisa baseada em gráficos e um atlas. Recursos semânticos, como a utilização de ontologias ou

tesauros, para indexar ou auxiliar na recuperação ou navegação, estão fora do escopo dessa proposta.

A pesquisa de Miguel (2009) analisa os principais problemas que afetam os SRI em IDE relacionados à gestão de ontologias terminológicas. O autor analisa e propõe uma alternativa para representação de modelos terminológicos (Figura 2.4).



Fonte: Miguel (2009)

Figura 2.4: Modelo alternativo para representação de um tesauro.

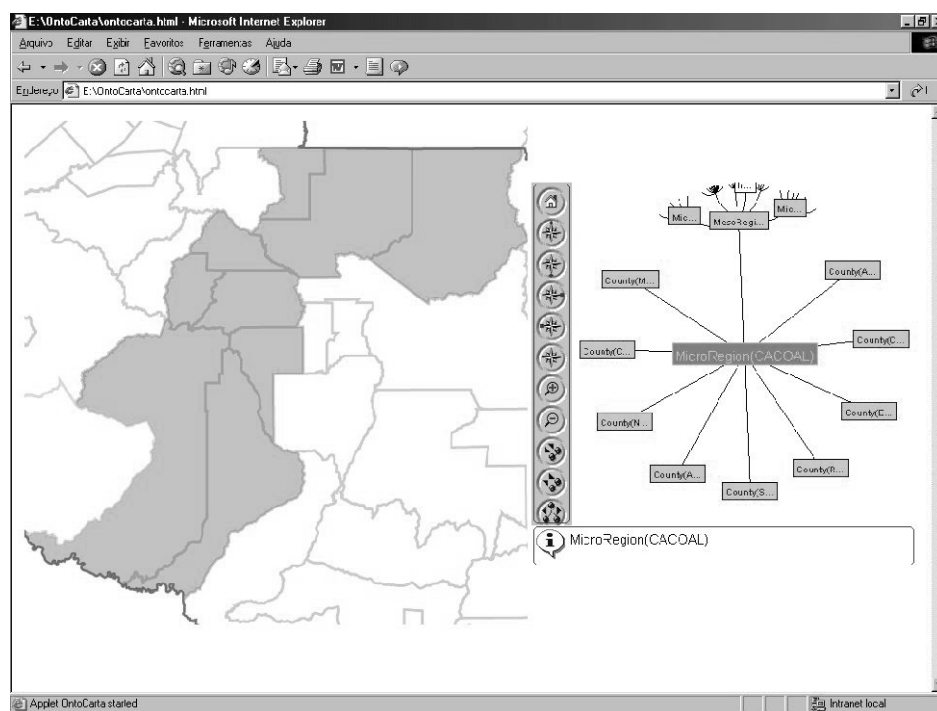
Nesta proposta, a classe “Conceito” é adicionada ao modelo relacionado com a classe “Termo”. A nova classe passa a ser o elemento central do modelo e recebe um identificador único. A integração do modelo terminológico à estrutura de uma IDE, os procedimentos para evolução do seu conteúdo e o acesso eficiente a estes modelos por *Web Services* são abordados e propostos como melhorias. Os processos, algoritmos, padrões, modelos e serviços desenvolvidos foram aplicados ao contexto de uma IDE em relação aos problemas detectados e o resultado foi promissor. A principal abordagem de Miguel (2009) é a integração de modelos terminológicos em IDE com o objetivo de simplificar a classificação dos recursos e melhorar a RI. Entretanto, o autor não chega a abordar uma forma de apresentar as soluções propostas aos usuários de uma IDE, por exemplo uma interface de consulta e pesquisa.

No Brasil, também existem grupos que desenvolvem trabalhos relacionados aos temas abordados. Pesquisadores que atuam no Laboratório de Sistema de Informação⁷ (LIS) do Instituto de Computação (IC), da Universidade Estadual de Campinas

⁷ <http://www.lis.ic.unicamp.br/>

(UNICAMP), vêm trabalhando com modelagem e aplicações que manipulam grandes e complexos repositórios de dados.

Dentre os trabalhos realizados no LIS do IC-UNICAMP, Daltio e Medeiros (2008) apresentam um serviço Web cujo objetivo é auxiliar na pesquisa em biodiversidade. Este cenário apresenta alguns desafios de pesquisa em Ciência da Computação, tanto no nível físico (por exemplo, diversidade de estruturas de dados) quanto conceitual (por exemplo, diversidade de perspectivas e de domínios de conhecimento). As autoras propõem o uso de ontologias como uma forma de se resolver problemas de heterogeneidade. Isso, entretanto, gera novos desafios de pesquisa relativos a projeto, gerenciamento e compartilhamento de ontologias. Outro trabalho é o *OntoCarta*, um sistema de navegação em mapas apoiada por ontologia (VENANCIO et al., 2003). Ao executar um *zoom in* em um mapa digital, geralmente, o usuário perde a noção de contexto. A navegação apoiada por ontologia visa solucionar esse problema. Uma ontologia relacionada aos conceitos e objetivos do território permite localizar o objeto em foco numa base de conhecimento representada por uma árvore hiperbólica (Figura 2.5).

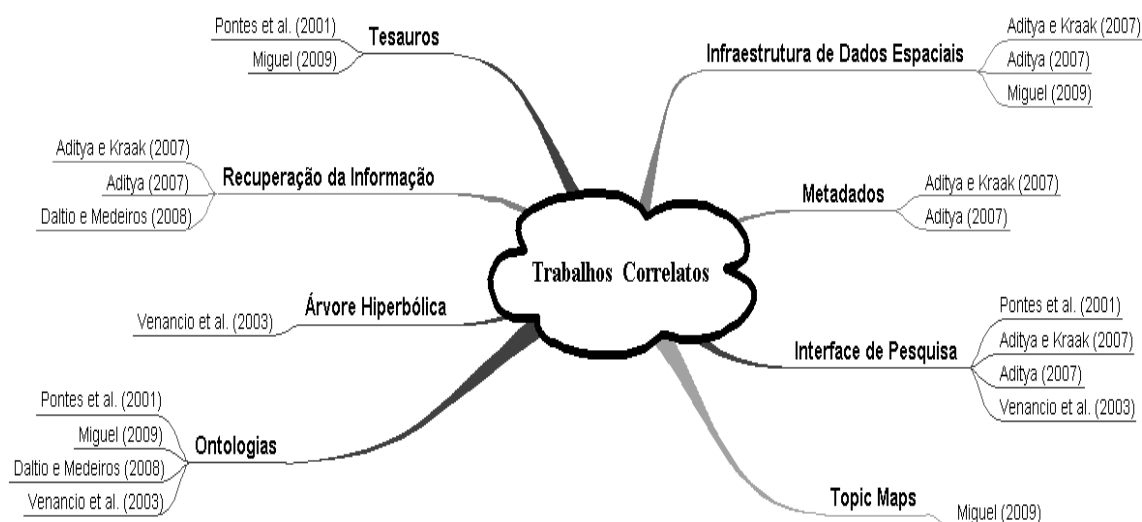


Fonte: Venancio et al. (2003)

Figura 2.5: Navegação em mapas dirigida por ontologia

Na literatura, foram encontrados vários trabalhos relacionados a RI e que utilizam tesauro e ontologia como instrumento de apoio. No entanto, não foram encontrados

trabalhos que abordam a RI em catálogos de metadados de um IDE combinados com mecanismos funcionais e intuitivos ao usuário. Um resumo das contribuições que influenciaram o desenvolvimento dessa dissertação é apresentado na forma de mapa mental (Figura 2.6). Esse mapa é o *Mind Map*, que utiliza a técnica proposta por Buzan (1996).



Fonte: Elaborado pelo Autor

Figura 2.6: Mapa mental dos trabalhos correlatos.

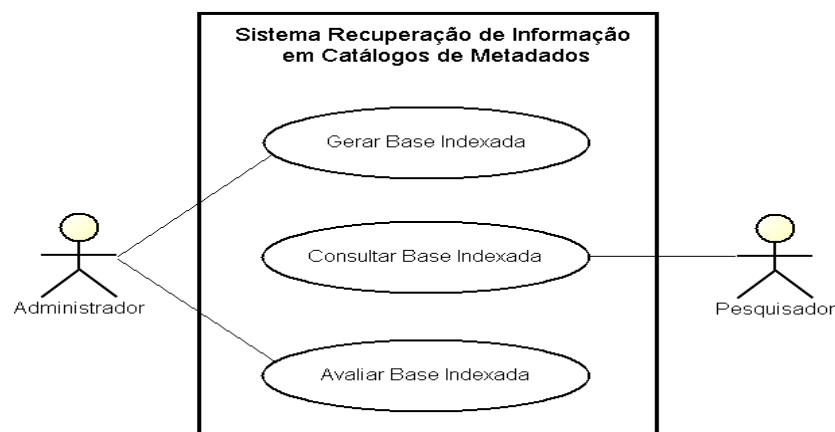
3 UMA ARQUITETURA PARA SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÃO EM CATÁLOGOS DE METADADOS

Este capítulo apresenta uma proposta de arquitetura para um sistema de recuperação de informação em catálogo de metadados de uma IDE. As Seções 3.1 e 3.2 apresentam, respectivamente, uma visão geral da arquitetura e os componentes do sistema. As Seções 3.3, 3.4 e 3.5 descrevem, em mais detalhes, os módulos que compõem o sistema. A seção 3.6 apresenta a estrutura e a comunicação entre as componentes da arquitetura do sistema.

3.1 Visão geral

O objetivo do sistema é prover uma alternativa para as dificuldades em RI em catálogos de metadados em uma IDE. Com isso, espera-se não apenas minimizar parte desses problemas, como também associar mais semântica às operações de consulta a catálogos de metadados. A especificação do sistema é voltada às necessidades da IDE de um projeto do contexto ambiental. Entretanto, essa especificação e o desenvolvimento do sistema têm sido conduzidos de maneira que ele possa ser aplicado em outras IDEs.

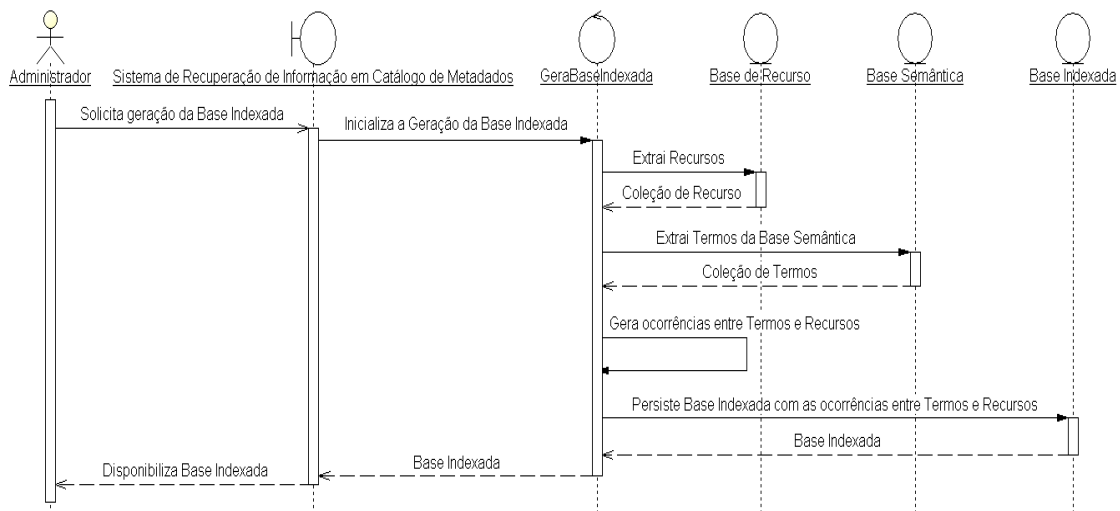
Para descrever o sistema, são utilizados diagramas da UML (*Unified Modeling Language*), como o diagrama de casos de uso, de sequência, de atividades, de componentes e de classes (Booch et al., 2005). Os casos de uso da Figura 3.1, em notação UML, definem os serviços fornecidos pelo sistema de recuperação de informação em catálogos de metadados. Os serviços disponíveis ao usuário baseiam-se em gerar, consultar e avaliar uma base de dados.



Fonte: Elaborado pelo Autor

Figura 3.1: Casos de uso do sistema de recuperação de informação em catálogos de metadados.

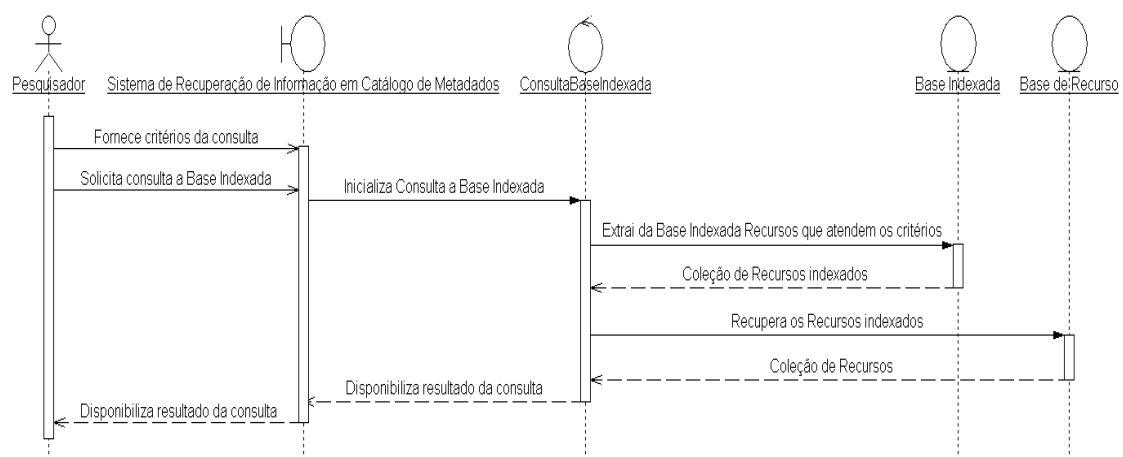
O caso de uso Gerar Base Indexada deve ser capaz de extrair termos e relacionamentos da base semântica (Tesauro), conteúdos da base de recurso (catálogo de metadados) e gerar a base indexada (*topic maps*) com termos, relacionamentos e ocorrências entre os termos e conteúdos do recurso. A Figura 3.2 ilustra o diagrama de sequência deste caso de uso.



Fonte: Elaborado pelo Autor

Figura 3.2: Diagrama de sequência do caso de uso Gerar Base Indexada.

O caso de uso Consultar Base Indexada deve ser capaz de extrair as ocorrências dos recursos da base indexada que atendem aos critérios da consulta e disponibilizar ao usuário os recursos recuperados na base de recurso. A Figura 3.3 mostra o diagrama de sequência deste caso de uso.

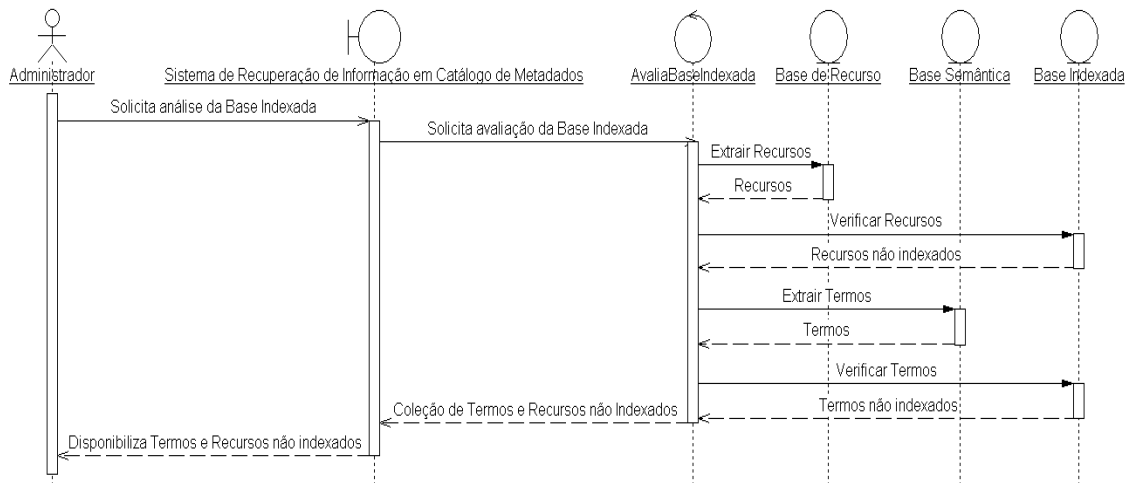


Fonte: Elaborado pelo Autor

Figura 3.3: Diagrama de sequência do caso de uso consultar base indexada.

O caso de uso Avaliar Base Indexada deve ser capaz de extrair recursos da base de recurso e termos da base semântica, bem como disponibilizar ao usuário uma lista dos

termos e recursos que não foram indexados na base. A Figura 3.4 mostra o diagrama de sequência deste caso de uso.

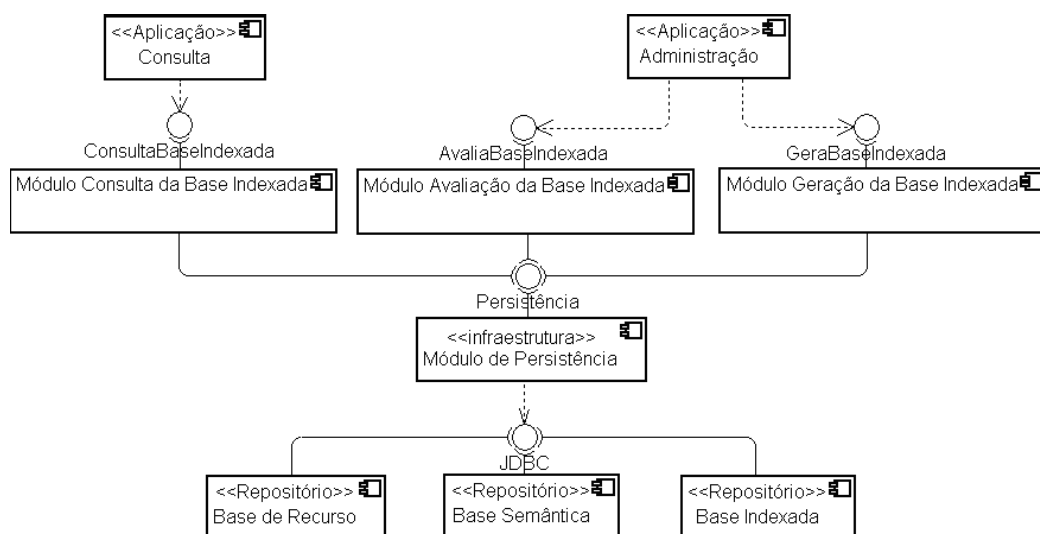


Fonte: Elaborado pelo Autor

Figura 3.4: Diagrama de sequência do caso de uso Avaliar Base Indexada.

3.2 Componentes do sistema

Para que os casos de uso definidos na seção anterior possam ser implementados, foram especificados os componentes da arquitetura do sistema de recuperação de informação em catálogo de metadados. O sistema é composto de três módulos que disponibilizam seus serviços por meio de classes de interface. Também fazem parte da arquitetura do sistema duas aplicações (uma para administração e outra de consulta), um módulo que controla a persistência e três repositórios de dados. Os componentes e seus relacionamentos estão ilustrados na Figura 3.5 e descritos a seguir.



Fonte: Elaborado pelo Autor

Figura 3.5: Diagrama de componentes do sistema.

Descrição dos componentes:

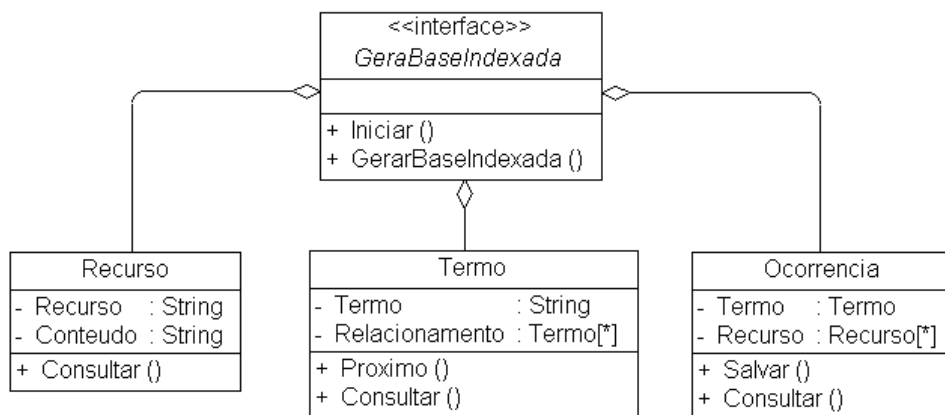
- Base de recurso: repositório que armazena os recursos coletados no catálogo de metadados da IDE. O recurso é composto pela identificação do metadado e parte de seu conteúdo. O conteúdo é formado por elementos (ex: resumo, palavras-chave) extraídos da estrutura do metadado, que geralmente estão armazenados em XML;
- Base semântica: repositório que armazena o conteúdo semântico de um tesauro. Sua estrutura permite armazenar a identificação do tesauro, seus termos, conceitos e relacionamentos. A estrutura foi desenvolvida seguindo as especificações ISO 2788 (ISO, 1986) e ISO 5964 (ISO, 1985) para construção de tesouros;
- Base indexada: repositório que armazena o resultado do processo de indexação entre as bases de recurso e semântica. Sua estrutura armazena os tópicos (termos do tesauro), suas associações (relacionamentos dos termos) e ocorrências (relação entre o termo do tesauro e a identificação do metadado). O resultado do processo de indexação segue as especificações do ISO 13250 (ISO, 2002) para construção de *topic maps*;
- Módulo geração da base indexada: módulo responsável por implementar uma interface com a aplicação de administração e gerenciar o processo de geração da base indexada como um todo;
- Módulo consulta da base indexada: responsável por implementar a interface com a aplicação de consulta e gerenciar o processo de consulta à base indexada;
- Módulo avaliação da base indexada: responsável por implementar uma interface com a aplicação de administração e gerenciar o processo de avaliação da base indexada. Esse processo faz uma análise da base integrada em relação às bases de recurso e semântica, gerando uma lista de quais metadados e termos não foram indexados;
- Módulo de persistência: módulo responsável por controlar a persistência e a recuperação de dados nos repositórios;
- Aplicação de administração: aplicação responsável por traduzir as funcionalidades dos módulos de geração e avaliação da base indexada para algo que o usuário consiga entender e assim interagir com o sistema;
- Aplicação de consulta: aplicação responsável por traduzir as funcionalidades do

módulo de consulta da base indexada para algo que o usuário consiga entender e assim interagir com o sistema. Essa aplicação apresenta a base indexada por meio de uma árvore hiperbólica, onde o usuário pode navegar, selecionar termos de seu interesse e, por fim, processar sua pesquisa.

A seguir, os módulos que compõem as principais funcionalidades propostas na arquitetura são descritos em mais detalhes. Uma visão da implementação dos repositórios e aplicações da arquitetura é apresentada no Capítulo 4.

3.3 Módulo Geração da Base Indexada

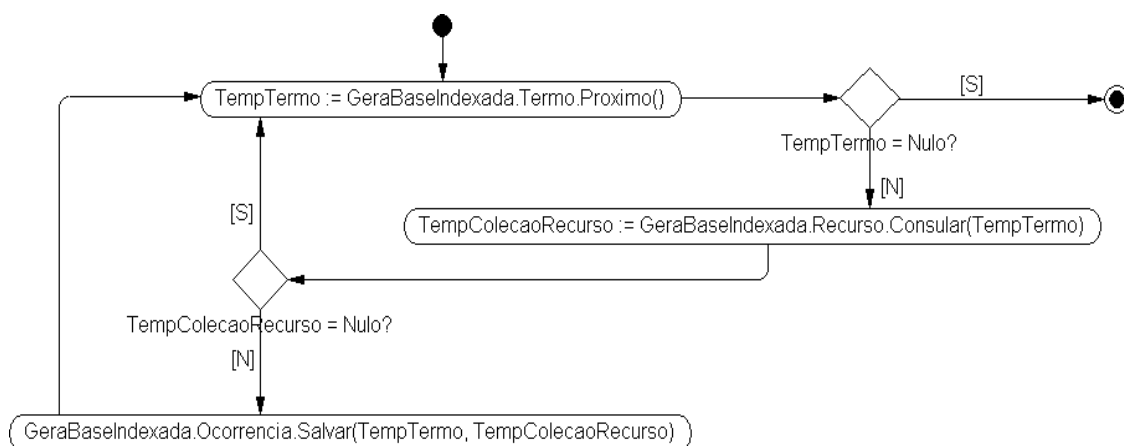
A Figura 3.6 ilustra o diagrama de classes do módulo Geração da Base Indexada. A interface *GeraBaseIndexada* define as funcionalidades do módulo, enquanto as demais são utilizadas para definir o tipo dos atributos desta classe. A classe *Recurso* define a estrutura dos recursos que são extraídos da base de recurso. A classe *Termo* define a estrutura dos termos que são extraídos da base semântica. A classe *Ocorrencia* define a estrutura dos recursos que serão persistidos na base indexada. A classe *GeraBaseIndexada* é responsável por todo o processo de geração da base indexada.



Fonte: Elaborado pelo Autor

Figura 3.6: Diagrama de classes do módulo Geração da Base Indexada.

Quando o usuário solicita a geração da base indexada, um objeto de uma classe que implementa a interface *GeraBaseIndexada* é instanciado e o método *Iniciar()* é invocado. O método *Iniciar()* é responsável por atualizar os atributos de *Recurso* e *Termo* com coleções recuperadas das bases de recurso e semântica. Em seguida, o método *GerarBaseIndexada()* é invocado. O método *GerarBaseIndexada()* gerencia o processo de indexação a partir das coleções que foram atualizadas, seguindo o fluxo de atividades ilustrado na Figura 3.7.

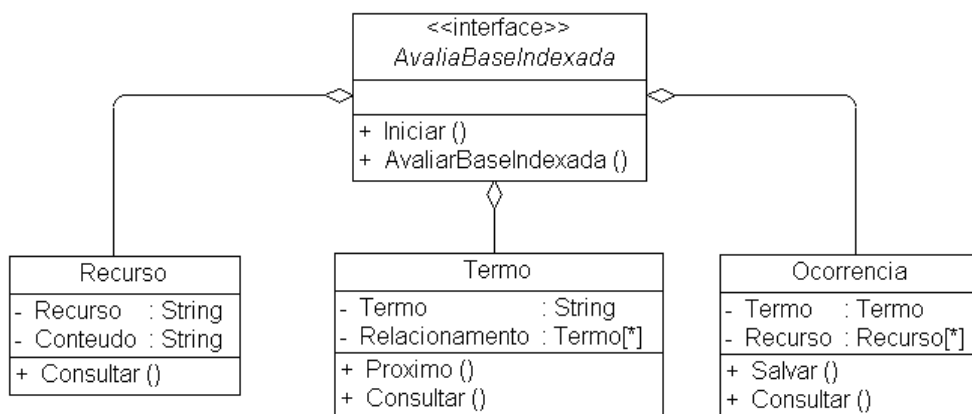


Fonte: Elaborado pelo Autor

Figura 3.7: Diagrama de atividades do método GerarBaseIndexada.

3.4 Módulo Avaliação da Base Indexada

A Figura 3.8 ilustra o diagrama de classes do módulo Avaliação da Base Indexada. A interface *AvaliaBaseIndexada* define as funcionalidades do módulo, enquanto as demais são utilizadas para definir o tipo dos atributos desta classe. A classe *Recurso* define a estrutura dos recursos que são extraídos da base de recurso. A classe *Termo* define a estrutura dos termos que são extraídos da base semântica. A classe *Ocorrencia* define a estrutura das ocorrências que são extraídas da base indexada. A classe *AvaliaBaseIndexada* é responsável por todo o processo de avaliação da base indexada.

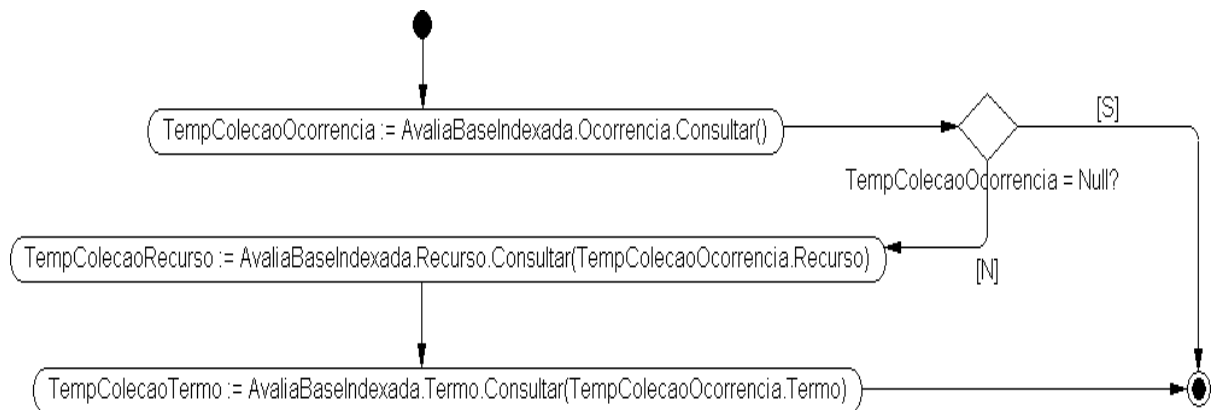


Fonte: Elaborado pelo Autor

Figura 3.8: Diagrama de classes do módulo Avaliação da Base Indexada.

Quando o usuário solicita a avaliação da base indexada, um objeto de uma classe que implementa a interface *AvaliaBaseIndexada* é instanciado e o método *Iniciar()* é invocado. O método *Iniciar()* é responsável por atualizar os atributos de *Recurso*, *Termo* e *Ocorrencia* com coleções recuperadas das bases de recurso, semântica e indexada. Em seguida, o método *AvaliaBaseIndexada()* é invocado. O método *AvaliaBaseIndexada()*

gerencia o processo de avaliação a partir das coleções que foram atualizadas, seguindo o fluxo de atividades ilustrado na Figura 3.9.

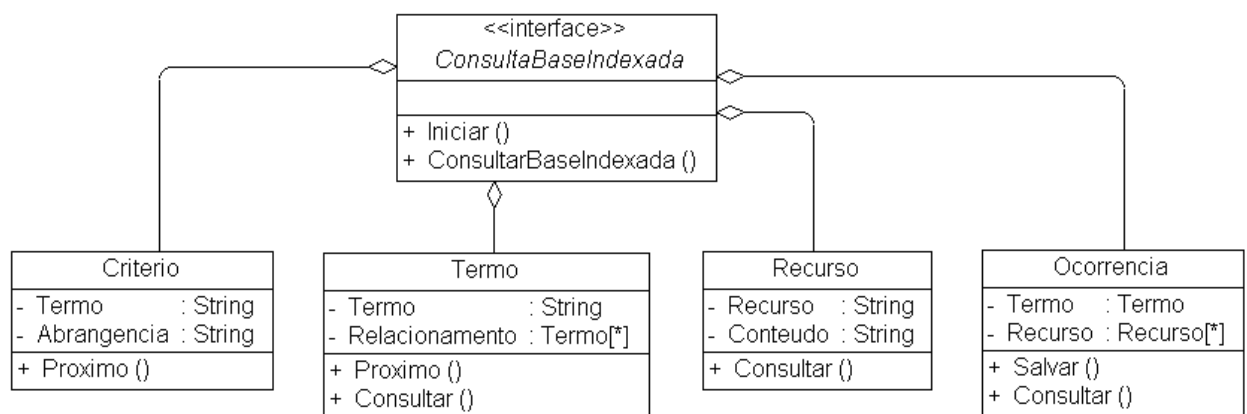


Fonte: Elaborado pelo Autor

Figura 3.9: Diagrama de atividades do método AvaliarBaseIndexada.

3.5 Módulo Consulta Base Indexada

A Figura 3.10 ilustra o diagrama de classes do módulo Consulta da Base Indexada. A interface ConsultaBaseIndexada define as funcionalidades do módulo, enquanto as demais são utilizadas para definir o tipo dos atributos desta classe. A classe Recurso define a estrutura dos recursos que são extraídos da base de recurso. A classe Termo define a estrutura dos termos que são extraídos da base semântica; a classe Ocorrencia, a estrutura dos recursos que são extraídos da base indexada; a classe Criterio, a estrutura dos critérios de consulta informados pelo usuário; e a classe ConsultaBaseIndexada é responsável por todo o processo de consulta a base indexada.

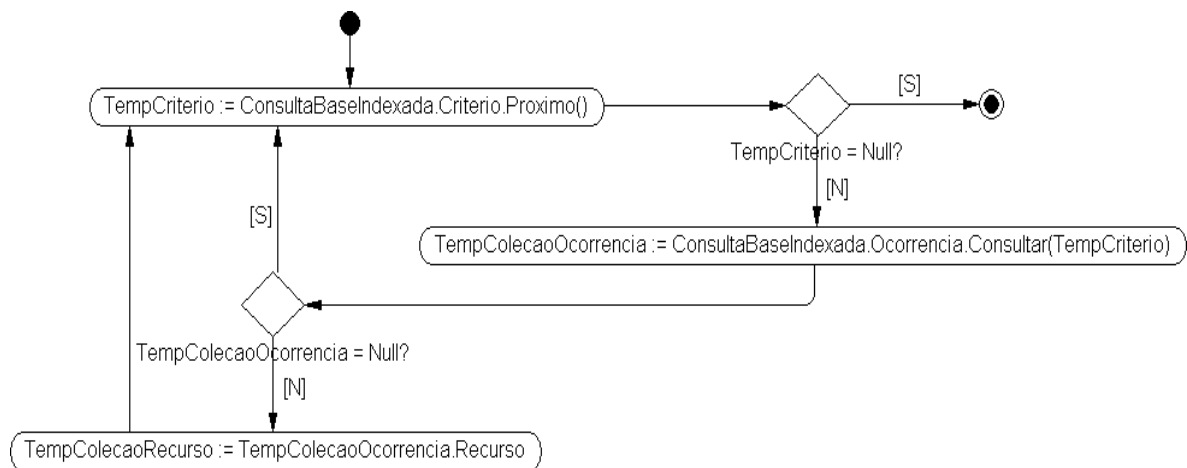


Fonte: Elaborado pelo Autor

Figura 3.10: Diagrama de classes do módulo consulta da base indexada.

Quando o usuário solicita uma consulta à base indexada, um objeto de uma classe que implementa a interface ConsultaBaseIndexada é instanciado e o método Iniciar() é

invocado. O método `Iniciar()` é responsável por atualizar os atributos de `Recurso`, `Termo`, `Ocorrencia` e `Criterio` com coleções recuperadas das bases de recurso, semântica, indexada e os critérios da pesquisa informados pelo usuário. Em seguida, o método `ConsultarBaseIndexada()` será invocado. Esse método gerencia todo o processo de consulta, seguindo o fluxo de atividades ilustrado na Figura 3.7.



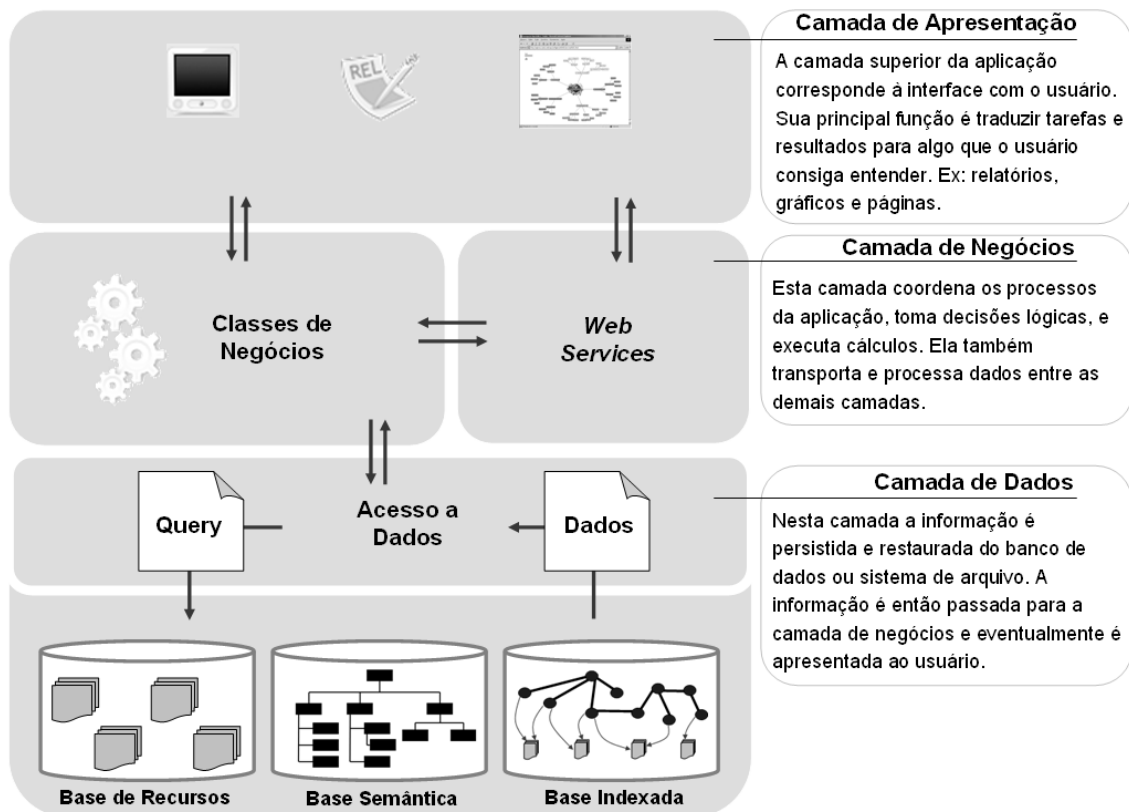
Fonte: Elaborado pelo Autor

Figura 3.11: Diagrama de atividades do método `ConsultarBaseIndexada`.

3.6 Estrutura da arquitetura

Com o aumento da complexidade das aplicações, torna-se fundamental a adoção de padrões que auxiliem no desenvolvimento de *softwares*. Em termos de estruturas para *softwares*, o padrão de arquitetura MVC (*Model-View-Controller*) fornece uma maneira de dividir as funcionalidades envolvidas na manutenção e apresentação dos dados de uma aplicação. O principal objetivo desse padrão é separar as estruturas da aplicação em três áreas: apresentação, lógica de negócios e dados.

A Figura 3.12 ilustra a estrutura da arquitetura proposta para o sistema, sendo composta por três camadas: Dados, Negócio e Apresentação. A Camada de Dados é composta por três repositórios: Base de Recurso, Base Semântica e Base Integrada. Os repositórios são acessados apenas pela camada de negócio. A camada de negócio é composta pelas classes de negócios e por *Web Services*, que permitem aplicações clientes interagirem com o sistema. Todas as funcionalidades do sistema são implementadas nas classes de negócios. A camada de apresentação traduz as funcionalidades das classes de negócios para algo que o usuário consiga entender e assim interagir com o sistema.



Fonte: Elaborado pelo Autor

Figura 3.12: Estrutura da arquitetura.

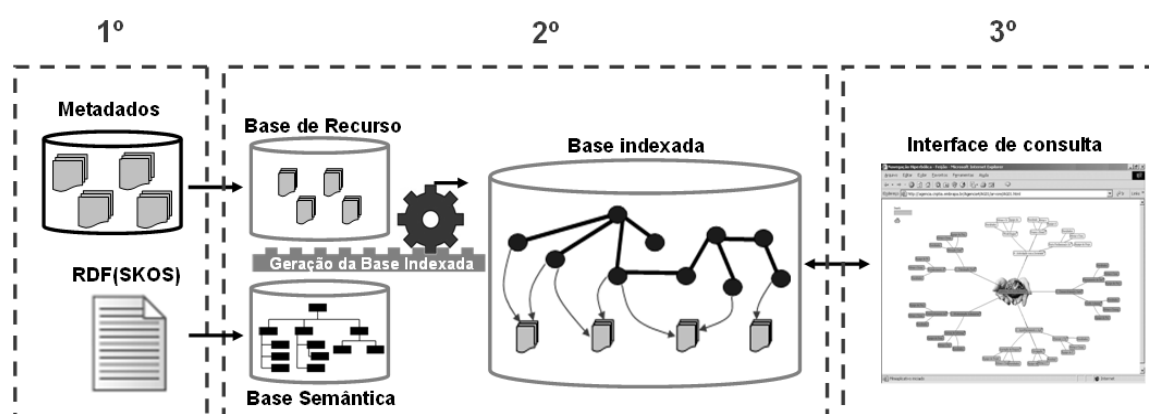
4 ASPECTOS DE IMPLEMENTAÇÃO

Este capítulo descreve os aspectos da implementação de um protótipo utilizando a arquitetura proposta no Capítulo 3. A Seção 4.1 apresenta uma visão geral do protótipo. As Seções 4.2 e 4.3 descrevem, em mais detalhes, o desenvolvimento do repositório e as funcionalidades do protótipo.

4.1 Visão geral

Para verificar a funcionalidade da arquitetura proposta no Capítulo 3, desenvolveu-se um protótipo de implementação, o qual é apresentado a seguir. O protótipo do SRI em catálogo de metadados foi implementado na linguagem Java⁸. O Sistema de Gerenciamento de Banco de Dados Relacional (SGBDR) PostgreSQL⁹ foi utilizado para armazenar parte dos dados necessários ao funcionamento do sistema. O acesso e a manipulação do conteúdo de documentos XML são providos pela API (*Application Programming Interface*) DOM (*Document Object Model*). Para a geração da árvore hiperbólica, foi utilizado o pacote Java Treebolic¹⁰.

O funcionamento do protótipo divide-se em três etapas, como ilustra a Figura 4.1. O primeiro passo é importar o catálogo de metadados e tesauro para as bases de recurso e semântica, respectivamente. Em seguida, executar a rotina de geração da base indexada e, por último, executar a aplicação de consulta à base indexada.



Fonte: Elaborado pelo Autor

Figura 4.1: Etapas de funcionamento do protótipo

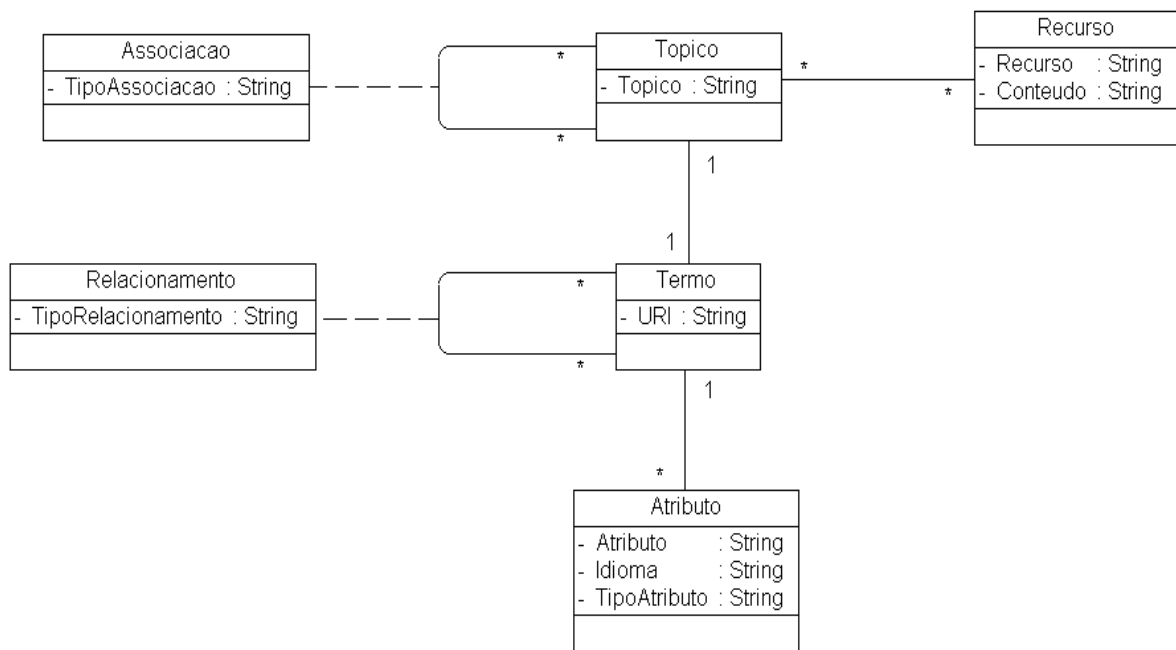
⁸ Java - <http://java.sun.com/>

⁹ PostgreSQL - <http://www.postgresql.org>

¹⁰ Treebolic - <http://treebolic.sourceforge.net>

4.2 Criação e carga do repositório

Para a realização dos testes do protótipo, foi implementada uma infraestrutura para a construção do repositório. O SGBDR PostgreSQL foi adotado na implementação da infraestrutura do repositório. As facilidades em armazenar, manipular e recuperar dados em um SGBDR justificam sua adoção em relação à opção implementar essa infraestrutura em arquivos semiestruturados. A Figura 4.2 apresenta o esquema conceitual do repositório em um diagrama de classe UML.



Fonte: Elaborado pelo Autor

Figura 4.2: Esquema conceitual do repositório.

A estrutura do modelo conceitual foi especificada em relação às necessidades de armazenamento identificadas nas três bases, as quais são descritas a seguir:

- Base de recurso: armazena o conteúdo dos elementos extraídos de metadados espaciais, estruturados em arquivos XML;
- Base semântica: armazena termos e relacionamentos extraídos de um tesauro, estruturado em um arquivo SKOS; e
- Base indexada: armazena os tópicos, associações e ocorrências que compõem a estrutura de um *topic maps*.

A relação entre as necessidades de armazenamento e a especificação do esquema conceitual é apresentada na forma de uma matriz de rastreabilidade, ilustrada pela Figura 4.3.

	Base de Recurso	Base Semântica	Base Indexada
Recurso	■	☐	■
Topico	☐	☐	■
Associacao	☐	☐	■
Termo	☐	■	☐
Relacionamento	☐	■	☐
Atributo	☐	■	☐

Fonte: Elaborado pelo Autor

Figura 4.3: Matriz de rastreabilidade: Classes X Bases.

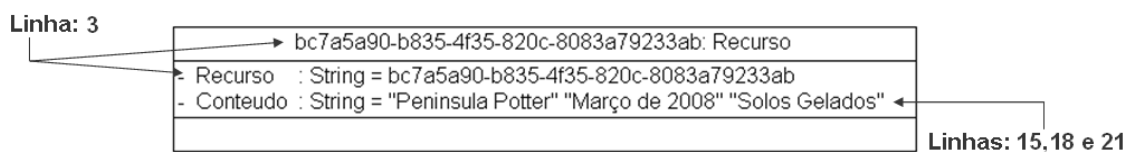
A rotina que extrai parte do conteúdo dos metadados, termos, atributos e relacionamentos de um tesauro foi implementada utilizando-se a API DOM. Essa API trata o documento XML como uma estrutura hierárquica de árvore, onde cada elemento da árvore é considerado um nó (*nodo*). Cada nó possui propriedades e métodos que podem ser aplicados ou manuseados. Do ponto de vista de uma linguagem orientada a objetos, como Java, todo tipo de nó possui uma interface correspondente. A API DOM oferece uma interface para acessar e manipular os objetos. Dessa forma, a API DOM é bastante apropriada para a importação de trechos de documentos XML para a estrutura do repositório.

Para facilitar seu desenvolvimento e tratar algumas características na lógica de leitura dos arquivos de metadados e tesauro, as rotinas de importação foram implementadas separadamente. A primeira delas, que importa os metadados, carrega um arquivo XML em memória e, em seguida, localiza o identificador do metadado e as suas palavras-chave. Essas duas informações serão persistidas no repositório pela classe Recurso, respectivamente nos atributos de Recurso e Conteúdo. Para simplificar os testes e o desenvolvimento da rotina que gera as ocorrências entre as classes Recurso e Topico, apenas as palavras-chave do metadado são utilizadas para compor o atributo Conteúdo da classe Recurso. Havendo a necessidade, esse atributo pode ser composto por outros elementos do arquivo XML, por exemplo, título e resumo.

Em uma nova versão deste protótipo, essa rotina pode indexar as ocorrências em outros elementos do arquivo XML, por exemplo título e resumo.

Para exemplificar o funcionamento da rotina que importa parte do conteúdo de um metadado, a Figura 4.4 ilustra um objeto da classe Recurso referente ao conteúdo do

trecho do arquivo XML da Figura 4.5.



Fonte: Elaborado pelo Autor

Figura 4.4: Objeto da classe Recurso.

```

1 <MD_Metadata>
2   <fileIdentifier>
3     <CharacterString>bc7a5a90-b835-4f35-820c-8083a79233ab</CharacterString>
4   </fileIdentifier>
5   <title>
6     <CharacterString>Perfis de Solo</CharacterString>
7   </title>
8   <abstract>
9     <CharacterString>Pontos de Perfis de coleta de amostras de solos para
10      analises fisico-quimicas durante Operantar n. 27
11      (2007/2008).</CharacterString>
12   </abstract>
13   <descriptiveKeywords>
14     <keyword>
15       <CharacterString>Peninsula Potter</CharacterString>
16     </keyword>
17     <keyword>
18       <CharacterString>Março de 2008</CharacterString>
19     </keyword>
20     <keyword>
21       <CharacterString>Solos Gelados</CharacterString>
22     </keyword>
23   </descriptiveKeywords>
24 </MD_Metadata>

```

Fonte: Elaborado pelo Autor

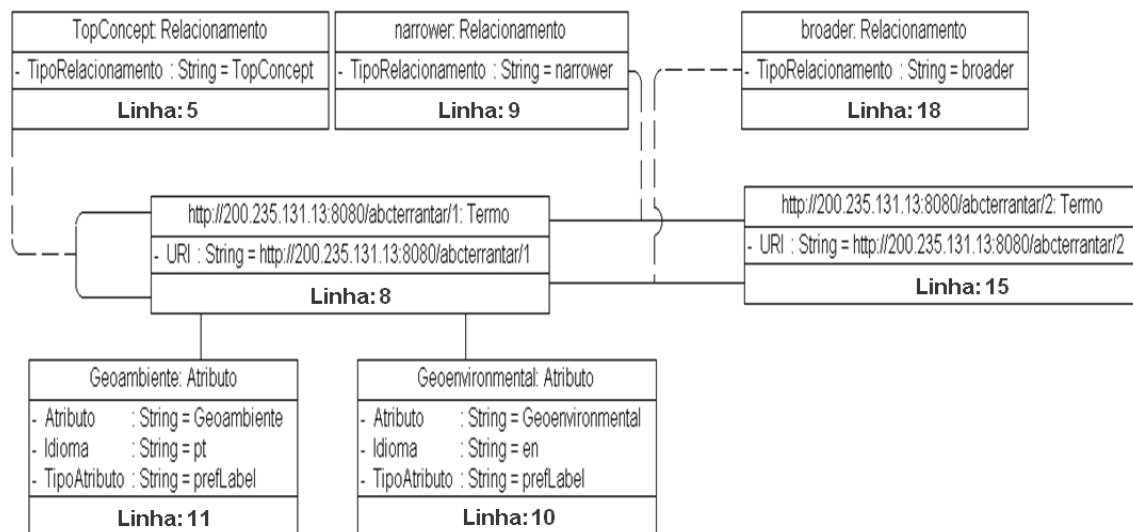
Figura 4.5: Trecho XML de um arquivo de metadado.

A segunda rotina, que importa os termos atributos e relacionamentos de um tesauro, carrega um arquivo RDF/SKOS, sendo a leitura da estrutura do tesauro mais complexa em relação ao metadado. Ao contrário da rotina que importa metadado e faz uma única passagem pelos nós, é preciso navegar entre os elementos e atributos do arquivo. O processo de leitura e persistência segue o roteiro:

1. Localizar os termos que possuem um relacionamento do tipo TopConcept;
2. Para cada termo do passo-1, localizar seus atributos e relacionamentos;
3. Para cada relacionamento do tipo Narrower do passo-2, localizar o termo correspondente, seus atributos e relacionamentos; e
4. Repetir o passo-3, substituindo os relacionamentos localizados na iteração atual, enquanto existir novos relacionamentos do tipo Narrower.

Os termos, atributos e relacionamentos são persistidos no repositório pelas classes Termo, Relacionamento e Atributo, ao final de cada passo. Para exemplificar o funcionamento da rotina que importa o conteúdo de tesauro, a Figura 4.6 ilustra os

objetos das classes Termo, Relacionamento e Atributo, referentes ao conteúdo do trecho do arquivo RDF/SKOS, da Figura 4.7.



Fonte: Elaborado pelo Autor

Figura 4.6: Objetos das classes Termo, Relacionamento e Atributo.

```

1 <rdf:RDF
2   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3   xmlns:skos="http://www.w3.org/2004/02/skos/core#" >
4   <rdf:Description rdf:about="http://iaaa.unizar.es/thesaurus/abcterrantar">
5     <skos:hasTopConcept rdf:resource="http://200.235.131.13:8080/abcterrantar/1"/>
6     <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#ConceptScheme"/>
7   </rdf:Description>
8   <rdf:Description rdf:about="http://200.235.131.13:8080/abcterrantar/1">
9     <skos:narrower rdf:resource="http://200.235.131.13:8080/abcterrantar/2"/>
10    <skos:prefLabel xml:lang="en">Geoenvironmental</skos:prefLabel>
11    <skos:prefLabel xml:lang="pt">Geoambiente</skos:prefLabel>
12    <skos:inScheme rdf:resource="http://iaaa.unizar.es/thesaurus/abcterrantar">
13      <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
14    </rdf:Description>
15    <rdf:Description rdf:about="http://200.235.131.13:8080/abcterrantar/2">
16      <skos:narrower rdf:resource="http://200.235.131.13:8080/abcterrantar/3"/>
17      <skos:prefLabel xml:lang="pt">Solo</skos:prefLabel>
18      <skos:broader rdf:resource="http://200.235.131.13:8080/abcterrantar/1"/>
19      <skos:prefLabel xml:lang="en">Soil</skos:prefLabel>
20      <skos:inScheme rdf:resource="http://iaaa.unizar.es/thesaurus/abcterrantar">
21        <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
22      </rdf:Description>

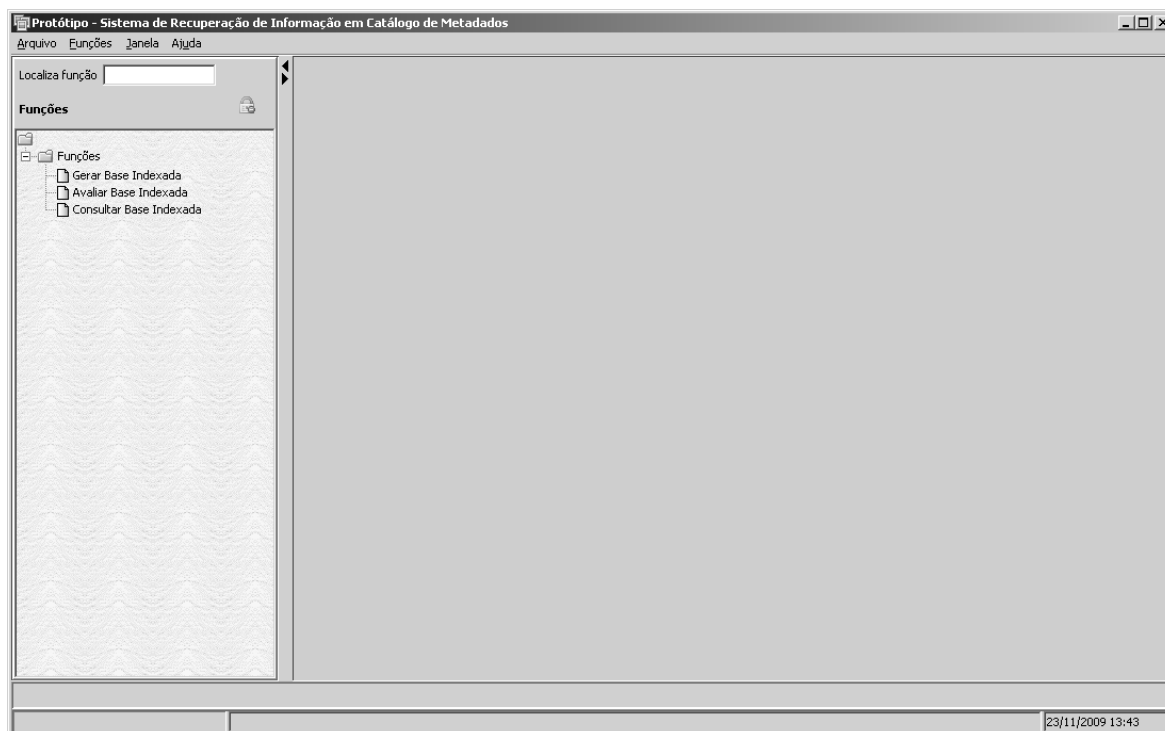
```

Fonte: Elaborado pelo Autor

Figura 4.7: Trecho RDF/SKOS de um arquivo de tesouro.

4.3 Funcionalidades do protótipo

A implementação das funcionalidades do protótipo seguiu as especificações do Capítulo 3. O protótipo é composto por cinco funções, sendo três referentes aos módulos da Figura 3.5 e duas referentes às rotinas especificadas na seção anterior. Por se tratar de um protótipo, todas as funções foram desenvolvidas na mesma aplicação. A Figura 4.8 ilustra a tela inicial do protótipo.



Fonte: Elaborado pelo Autor

Figura 4.8: Tela inicial do protótipo.

A tela inicial permite ao usuário gerar, avaliar e consultar a base indexada. O principal objetivo do protótipo é avaliar a proposta desta dissertação, que busca uma forma alternativa para a pesquisa em catálogos de metadados, sendo a fonte para essa pesquisa uma base indexada que segue a estrutura de *topic maps*. Tomou-se o cuidado em criar mecanismos que permitam não só a geração dessa base, mas também sua avaliação. O algoritmo para a geração da base indexada executa uma busca exaustiva na localização de ocorrências entre os tópicos e recursos existentes no repositório. Esse algoritmo possui uma implementação simples e sempre encontrará uma solução, se ela existir. Caso o conteúdo da Base de Recurso passe a ser composto por outros elementos do metadado, pode ser interessante analisar a utilização de algoritmos mais refinados.

As bases de recurso e semântica sofrem constantes alterações; este fato se deve à criação de novos metadados. Se acontecer qualquer alteração em alguma dessas bases, será necessário regerar a base indexada. A Figura 4.9 ilustra os recursos disponíveis para importar e atualizar a base de recurso e semântica, além de permitir a geração da base indexada.



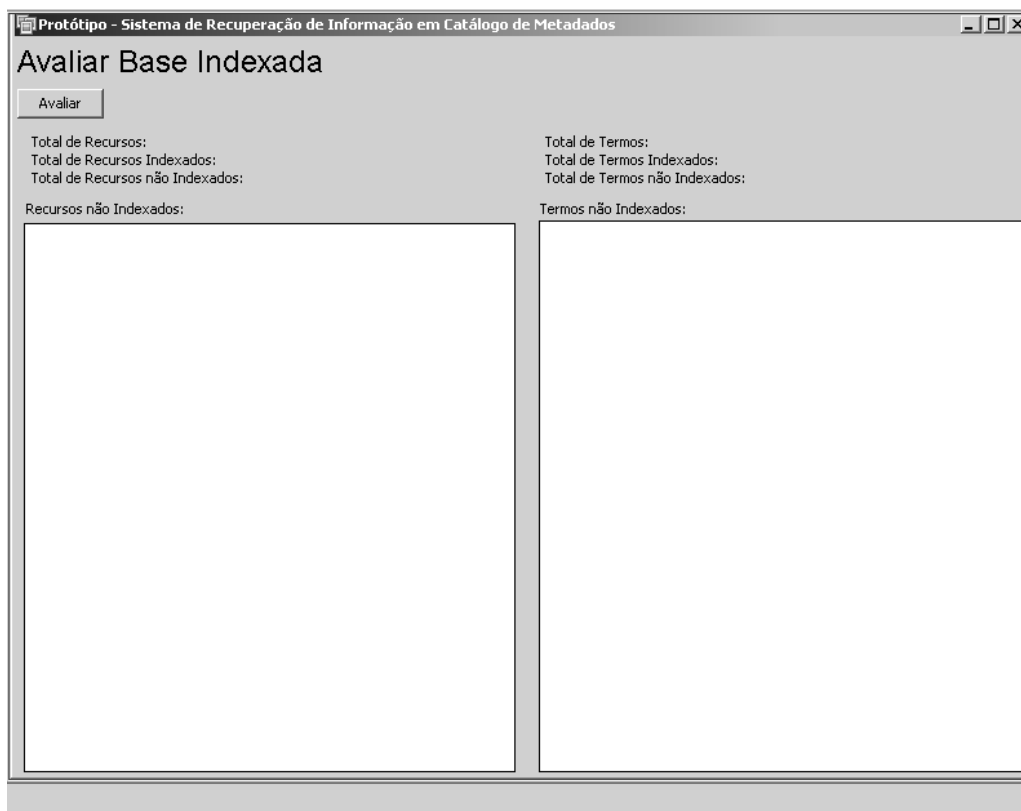
Fonte: Elaborado pelo Autor

Figura 4.9: Tela gerar base indexada do protótipo.

A avaliação da base indexada tem como objetivo quantificar os recursos e termos existentes no repositório. Ao final de sua execução, algumas informações são exibidas ao usuário:

- Total de recursos: quantidade de recursos existentes no repositório;
- Total de recursos indexados: quantidade de recursos que foram indexados;
- Total de recursos não indexados: quantidade de recursos que não foram indexados;
- Recursos não Indexados: listagem dos recursos que não foram indexados;
- Total de termos: quantidade de termos existentes no repositório;
- Total de termos indexados: quantidade de termos que foram indexados;
- Total de termos não indexados: quantidade de termos que não foram indexados;
- e
- Termos não indexados: listagem dos termos que não foram indexados.

Novos indicadores e outros recursos que auxiliem o usuário na avaliação da base indexada são sugeridos como trabalhos futuros no Capítulo 6. A Figura 4.10 ilustra os recursos disponíveis na avaliação da base indexada.



Fonte: Elaborado pelo Autor

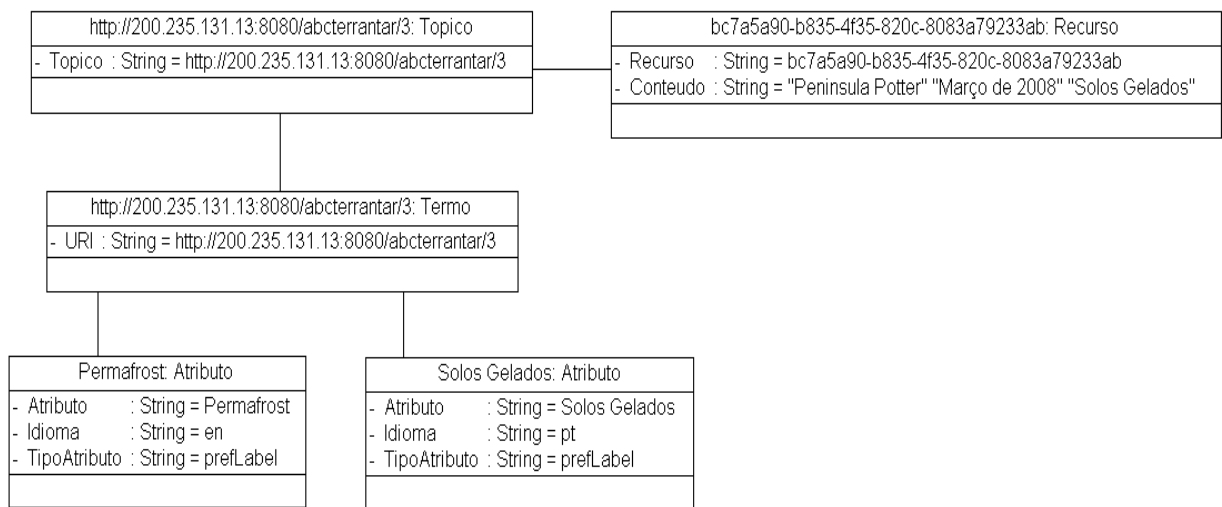
Figura 4.10: Tela avaliar base indexada do protótipo.

A interface de consulta desenvolvida tem como foco principal a semântica, ou seja, o entendimento do contexto na pesquisa é priorizado. A interface de pesquisa deve permitir ao usuário seguir uma linha de raciocínio em sua pesquisa, apoiado por uma árvore hiperbólica que busca representar o contexto das informações disponíveis. O usuário pode selecionar os termos e abrangência desejados e processar sua consulta.

A função que implementa a consulta na base indexada leva em consideração os critérios informados pelo usuário, que são a identificação do tópico e o tipo de associação. A primeira versão do protótipo suporta três tipos: Genérico (*Broader*), Específico (*Narrower*) e Relacionado (*Related*). Esses tipos caracterizam as associações entre os tópicos e recursos da base indexada e segue a semântica dos relacionamentos da estrutura do tesauro apresentado no Capítulo 2.

Para exemplificar o funcionamento da função de consultar a base indexada, a Figura 4.11 ilustra um conjunto de objetos das classes Tópico, Recurso, Termo e Atributo, referente à ocorrência entre um tópico e recurso. A ligação entre as classes Tópico e Termo indica que um tópico está relacionado ao termo e que este é composto por atributos que podem estar descritos em vários idiomas. A característica *multilíngue* foi levada em consideração no desenvolvimento da função de consulta, assim o usuário não

precisa informar o idioma do critério. Uma consulta que tenha o critério “Solos Gelados” iria recuperar recursos que tenham as ocorrências de “Solos Gelados” ou “*Permafrost*”.

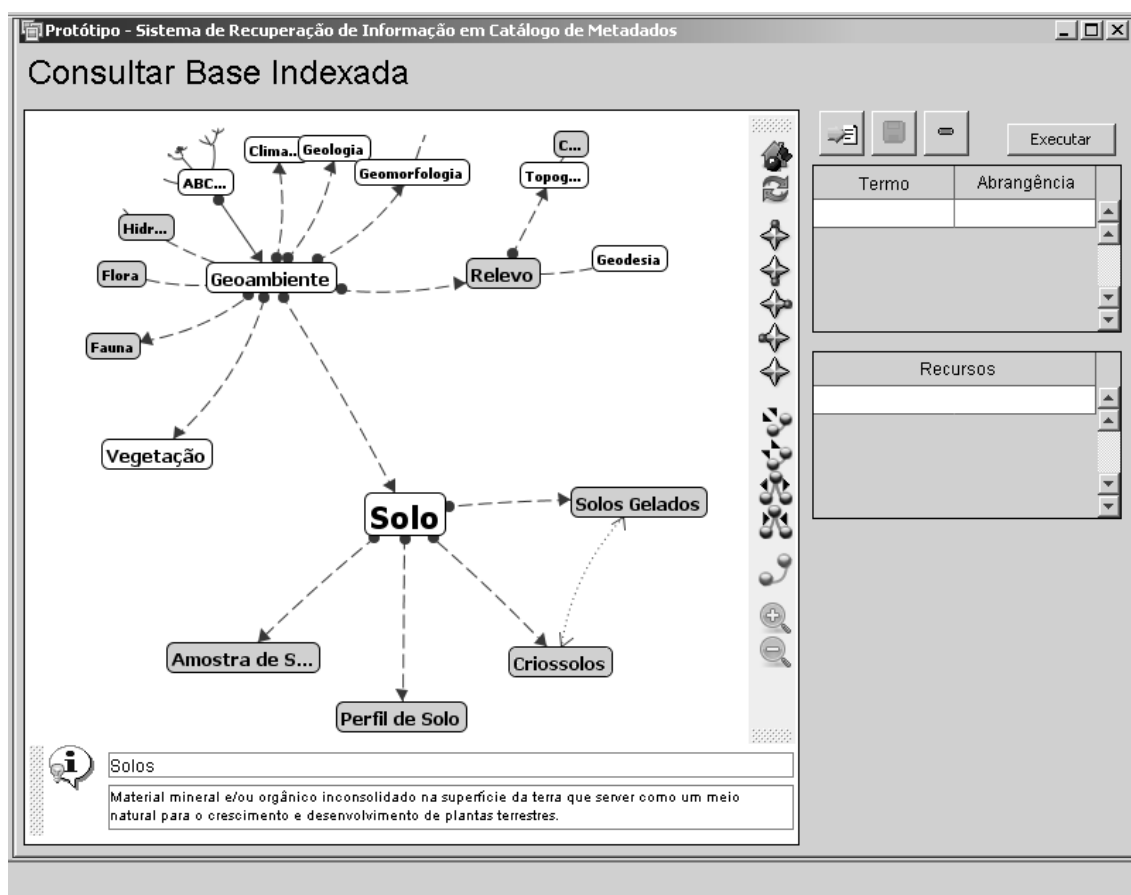


Fonte: Elaborado pelo Autor

Figura 4.11: Objetos das classes Termo, Relacionamento e Atributo.

A biblioteca Treebolic foi utilizada na geração da árvore hiperbólica. Os recursos dessa biblioteca permitiram combinar, destacar e exibir as informações do tesauro e *topic maps* na mesma interface visual. A primeira etapa da geração da árvore hiperbólica carrega os termos, notas de escopo e relacionamentos do tesauro. Os recursos da biblioteca permitem, por exemplo, alterar cada propriedade de um nó ou da ligação entre os nós. Uma importante informação a ser exibida ao usuário é a nota de escopo do termo do tesauro. A interface exibe essa informação quando o usuário posiciona o *mouse* sobre o termo. Outro recurso utilizado na primeira versão do protótipo foi diferenciar a ligação entre os nós da árvore. Essa funcionalidade pode ser desativada, caso dificulte a interpretação do usuário.

Após carregar todas as informações do tesauro, a função carrega os tópicos da base indexada. Os termos correspondentes a um tópico têm sua cor alterada (ex: “Perfil de Solo” e “Solos Gelados”). Essa alteração indica de forma visual quais termos têm recursos indexados, característica que poderá evitar consultas desnecessárias. A Figura 4.12 ilustra a interface de consulta do protótipo.



Fonte: Elaborado pelo Autor

Figura 4.12: Tela Consultar Base Indexada do protótipo.

5 ESTUDO DE CASO

Este capítulo apresenta o projeto TERRANTAR¹¹ (Seção 5.1), âmbito do estudo de caso desta dissertação. Em seguida, é descrito o desenvolvimento da IDE do projeto TERRANTAR (Seção 5.2). O processo de criação e evolução da base semântica é apresentado na Seção 5.3. A Seção 5.4 descreve os testes e avaliações do protótipo desenvolvido no Capítulo 4, no contexto da IDE do estudo de caso.

5.1 TERRANTAR - Ecossistemas terrestres da antártica

O TERRANTAR é um dos sete grupos que compõem o Instituto Nacional de Ciência e Tecnologia da *Criosfera*, o qual tem como missão implementar um programa nacional de pesquisas sobre as massas de gelo e neve no planeta. Cabe ao TERRANTAR articular e agregar os principais grupos brasileiros de pesquisa em criossolos (camada de solo que permanece abaixo de 0 °C por dois anos consecutivos), solos e ecossistemas terrestres, em suas diferentes áreas do conhecimento.

O objetivo do TERRANTAR é a articulação nacional e internacional de projetos de pesquisa visando à caracterização dos ecossistemas terrestres da antártica. A gerência e o compartilhamento dos resultados dessas pesquisas são alguns dos objetivos específicos do projeto. Esses resultados/dados estão registrados e armazenados em diversos formatos, como dados geográficos vetoriais ou matriciais, teses, artigos, planilhas e fotos. A primeira coleção de dados que está sendo catalogada e disponibilizada é resultado de projetos sediados no Departamento de Solos da Universidade Federal de Viçosa (UFV).

A primeira etapa do estudo de caso consistiu em definir o padrão de metadados geográficos a ser utilizado, visto que, embora a coleção de dados esteja sendo obtida por meio de projetos nacionais, estes dados poderão servir para pesquisadores de outros países. O padrão escolhido foi o ISO 19115, por ser um padrão internacional, ao contrário de padrões de uso particular, como o FGDC americano.

Para que a troca de informações seja eficiente, é necessário saber quais são os dados disponíveis e conhecer sua qualidade, fator contemplado pelos metadados. Entretanto, muitos usuários desconhecem as relações entre esses dados, o que limita sua potencialidade de uso e dificulta o intercâmbio. O desenvolvimento da IDE do projeto TERRANTAR levou em consideração os fatores tecnológicos e semânticos na

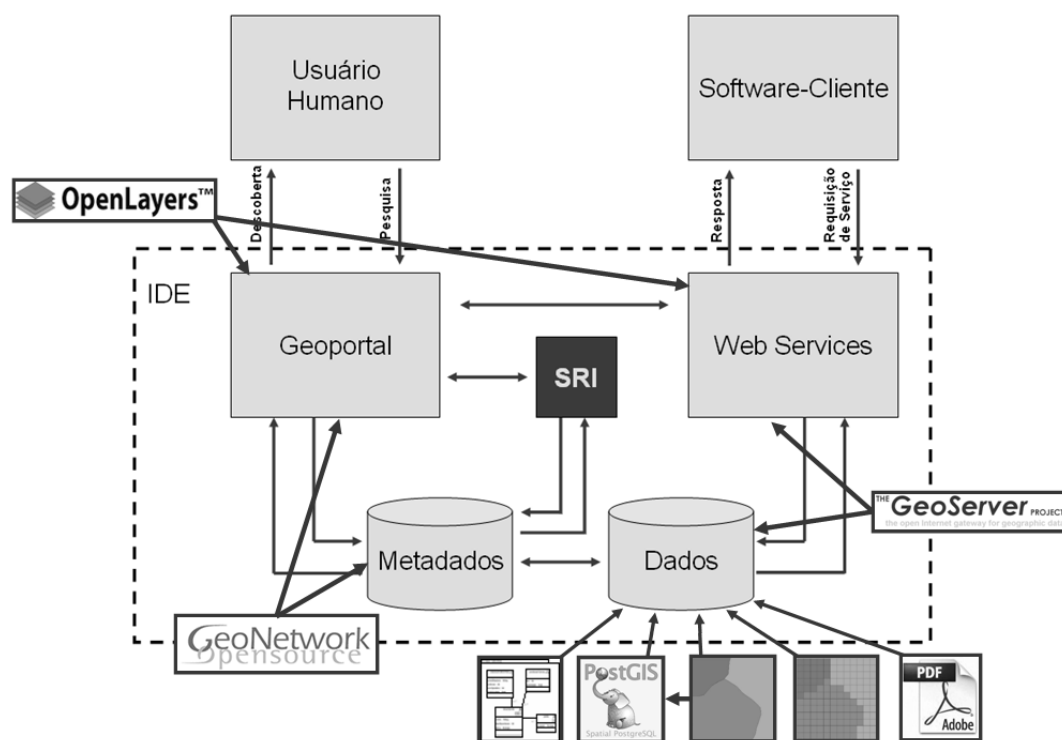
¹¹ TERRANTAR – Ecossistemas terrestres da antártica - <http://www.terrantar.com.br>

interoperabilidade dos dados.

5.2 Infraestrutura de dados espaciais do TERRANTAR

O principal objetivo da IDE TERRANTAR é apoiar o compartilhamento de dados geoespaciais. Contudo, a compatibilidade tecnológica é sempre uma preocupação. Uma das iniciativas mais recentes de promover interoperabilidade foi coordenada pelo *Open Geospatial Consortium* (OGC), o qual tem proposto padrões para o uso de serviços no compartilhamento de informação geográfica.

A Figura 5.1 ilustra a estrutura da IDE do projeto TERRANTAR enriquecida com um SRI de metadados desenvolvido com arquitetura proposta no Capítulo 3. A figura também apresenta as ferramentas computacionais *open source* adotadas na implantação da IDE e seu relacionamento com elementos dessa estrutura. Por fim, pode-se observar que os dados espaciais são armazenados em diversos formatos.



Fonte: Adaptado de Davis e Alves (2005)

Figura 5.1: IDE do projeto TERRANTAR.

5.2.1 Ferramentas computacionais

Diante das necessidades iniciais levantadas no projeto, um estudo de viabilidade técnica foi direcionado aos padrões, métodos, conceitos e às tecnologias existentes que pudessem embasar o seu desenvolvimento. O enfoque do estudo foi a busca de

ferramentas computacionais para otimizar e facilitar o desenvolvimento da IDE.

Algumas ferramentas (*open source*) foram selecionadas para implementação da IDE, as principais são: OpenLayer¹², GeoServer¹³, PostgreSQL e Geonetwork¹⁴. Essas ferramentas computacionais são detalhadas a seguir.

O Geonetwork é um sistema de gerenciamento de metadados por meio da Web. É muito utilizado para gestão de grandes volumes de dados geográficos, permitindo a organização destes através de padrões de metadados geográficos. Dentre os principais recursos e características, estão:

- Controle de acesso personalizado;
- Permissão para o armazenamento de metadados em diversos SGBDRs: MySQL, SQLServer, PostgreSQL e Oracle.
- Coleta programada e sincronização dos metadados distribuídos entre catálogos;
- *Uploading e downloading* de dados, documentos e outros conteúdos; e
- Suporte aos padrões de metadados ISO 19115:2003, FGDC e Dublin Core.

O PostgreSQL é um sistema gerenciador de banco de dados objeto relacional com suporte para tipos de dados geodésicos e geométricos. Este recurso é implementado em sua extensão espacial nomeada PostGIS. Dentre os seus principais recursos e características, estão:

- Suporte para índices espaciais GiST e R-Tree;
- Funções para análise básica e processamento de objetos geométricos.

O OpenLayer é uma biblioteca *JavaScript* que permite exibir mapas geográficos dinâmicos em uma página Web. Destacam-se nesta ferramenta os seguintes recursos e características:

- Armazena as informações no *cache* do navegador e atualiza apenas a parte do mapa que sofreu alguma alteração; e
- Permite utilizar funções básicas de navegação em mapas, por exemplo *pan*, *zoom in* e *zoom out*.

O GeoServer é um servidor de *Web Services* de padrões abertos que implementa, principalmente, as especificações OGC. Entre os principais recursos e características, estão:

¹² OpenLayer - <http://www.openlayers.org>

¹³ GeoServer - <http://www.geoserver.org>

¹⁴ Geonetwork - <http://geonetwork-opensource.org>

- Implementa os padrões de *Web Services* do OGC: WFS¹⁵, WFS-T¹⁶, WCS¹⁷ e WMS¹⁸; e
- Permite a integração com os SGBDR: MySQL, SQLServer, PostgreSQL e Oracle.

5.2.2 Funcionalidades e limitações

A catalogação e acesso aos metadados são providos pelo Geonetwork. O repositório de metadados é mantido pelo SGBDR PostgreSQL. A implementação de *Web Serviços* é provida pelo GeoServer. Essa ferramenta disponibiliza os dados geográficos em vários formatos, proporcionando interoperabilidade aos usuários. Estes serviços que permitem a interoperabilidade são descritos a seguir:

- Serviços WFS e WFS-T: provêm uma interface para inserir, selecionar, atualizar e remover feições geográficas, permitindo o seu processamento em aplicações clientes, como é o caso da edição vetorial em SIG;
- Serviço WCS: permite acesso a dados geográficos em formato matricial (*coverages*), para que possam ser processados;
- Serviço de WMS: produz mapas para a Web a partir de dados geográficos vetoriais e matriciais, de diferentes fontes.

Finalmente, o OpenLayer permite que os usuários visualizem e analisem os dados geográficos em mapas interativos sem a necessidade de baixá-los. Os mapas são gerados utilizando-se o serviço de WMS disponibilizado pelo GeoServer.

O aplicativo Geonetwork fornece uma solução para o portal de comunicação com os usuários. O Portal possui uma interface ilustrada pela Figura 5.2, que segue uma linha tradicional na recuperação de metadados. Segundo Hochmair (2005), como foi descrito no Capítulo 2, essa interface pode apresentar algumas dificuldades, especialmente para os usuários inexperientes, que podem não saber quais palavras-chave usar, como preencher adequadamente o formulário de consulta ou definir quantos critérios serão utilizados. Essa abordagem pode gerar um alto custo computacional na pesquisa, visto que geralmente os metadados estão armazenados em formatos semiestruturados (ex.: formato XML).

¹⁵ *Web Feature Service* (WFS): serviço para acesso e manipulação de dados geográficos

¹⁶ *Web Feature Service Transacional* (WFS-T): permite a alteração do dado

¹⁷ *Web Coverage Service* (WCS): serviço para acesso a dados geográficos do tipo *Coverage*

¹⁸ *Web Map Service* (WMS): serviço para produção de mapas.

Universidade Federal de Viçosa

Português-Brasil | English

TERRANTAR
ECO SISTEMAS TERRESTRES DA ANTÁRTICA

Nome de usuário Senha Entrar

BUSCAR

O QUE?

Título

Resumo

Palavras-Chave

Precisão da Busca

Precisa > > > ☐ ☒ ☐ ☐ ☐ > > > Imprecisa

QUANDO?

☒ Qualquer período

☐ de Para

Restringir a

Catálogo

Categorias

ONDE?

lat (min)

long (min)

long (max)

lat (max)

Fechar opções avançadas Reiniciar **Buscar**

Fonte: Elaborado pelo Autor

Figura 5.2: Interface de pesquisa da IDE do projeto TERRANTAR.

5.3 Criação e evolução da base semântica

A criação do tesauro (ABC do TERRANTAR) seguiu o método de garantia literária usado por Hulme (1950) apud Dodebei (2002). Nesta fase, foram definidos os principais temas/categorias do tesauro. Os primeiros termos foram registrados no tesauro. A Figura 5.3 ilustra o termo Solo, sua nota de escopo, seus termos genéricos e específicos na interface da ferramenta Tematres.

ABC do Terrantar

Buscar

Lista Hierarquica
Lista Alfabética
Sobre...
Minha conta

Solo

Data de criação: 24-Nov-2009
Termo aceito: 24-Nov-2009

► INÍCIO ► Geoambiente ► Solo

Nota de escopo

Material mineral e/ou orgânico inconsolidado na superfície da terra que server como um meio natural para o crescimento e desenvolvimento de plantas terrestres.

TG Geoambiente

Solo

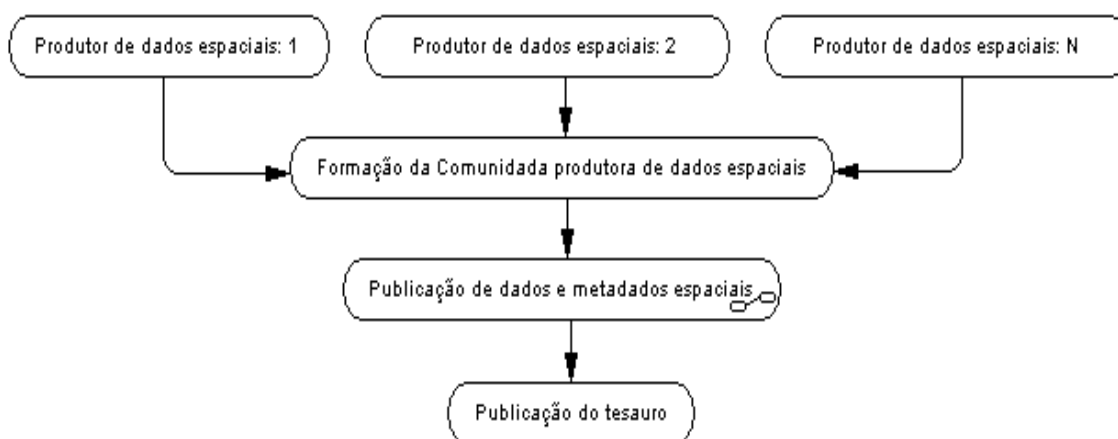
TE2 Amostra de Solo
TE2 Criossolos
TE2 Perfil de Solo
TE2 Solos Gelados

EQ Soil (ABC of Terrantar / en)

Fonte: Elaborado pelo Autor

Figura 5.3: Interface da ferramenta Tematres.

A evolução da base semântica está apoiada nas premissas da inteligência coletiva, em conjunto com um procedimento especificado para auxiliar os usuários produtores de dados espaciais da IDE do projeto TERRANTAR. O esquema geral do método utilizado é apresentado na Figura 5.4.



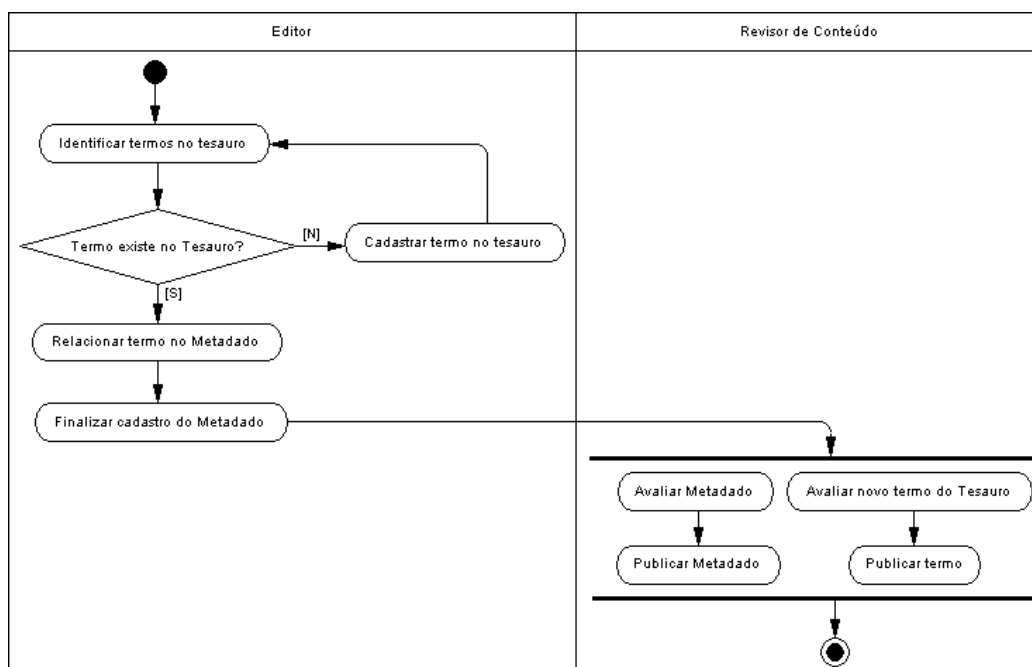
Fonte: Elaborado pelo Autor

Figura 5.4: Procedimento utilizado para evolução da base semântica.

A formação e criação do tesauro são sustentadas por um sistema de gestão de linguagens documentárias e sua evolução é influenciada pela interação dos usuários. Essa colaboração da “comunidade” segue o método de Endosso do Usuário, usado por Lancaster (1972) apud Dodebei (2002). A produção, coleta, avaliação e compartilhamento de informações são papéis assumidos pelos usuários da IDE. O procedimento utilizado é composto por três etapas, que são apresentadas a seguir:

- Formação da comunidade produtora de dados espaciais: Os elementos básicos de uma comunidade são as pessoas (por exemplo, membros, líderes, colaboradores), um contexto e seus objetivos. Nesta etapa, o usuário é informado das regras e normas a serem seguidas. O usuário é cadastrado e recebe um *login* e senha, de acordo com o papel (Administrador, Revisor de Conteúdo e Editor) a ser exercido por ele na comunidade;
- Publicação de dados e metadados espaciais: Com o uso de uma IDE, os “Editores” da comunidade podem publicar seus dados espaciais, juntamente com os metadados, que devem conter termos identificados no tesauro da IDE. Os dados e metadados são validados pelo “Revisor de Conteúdo”, que vai analisar todas as informações postadas pelos “Editores”. Por fim, o material liberado pelo “Revisor de Conteúdo” fica disponível à comunidade na IDE. Um esquema da etapa da publicação é apresentado no diagrama da Figura 5.5; e

- Publicação do tesauro: O tesauro publicado pode ser acessado pelo usuário da IDE na busca de conhecimento semântico. Em outro momento, essa relação entre os termos do tesauro e metadados permite a geração de *topic maps*, que será a base do SRI no catálogo de metadados da IDE.



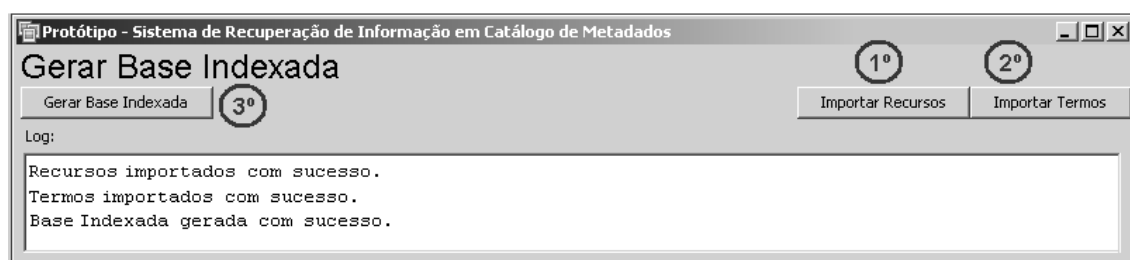
Fonte: Elaborado pelo Autor

Figura 5.5: Diagrama das atividades da etapa de publicação de dados e metadados.

As atividades da etapa de publicação de dados e metadados são ilustradas pela Figura 5.5. Por definição da equipe do projeto TERRANTAR, todo cadastro de metadado deve consumir termos existentes no tesauro da IDE. Caso o termo não exista, ele deve ser registrado no tesauro para que possa ser analisado e publicado. Essa definição possibilitou controlar os termos utilizados nas descrições dos metadados, fator que contribuiu para a geração da base indexada e eficiência no sistema de recuperação.

5.4 Testes do protótipo

Os testes realizados neste estudo de caso resolvem problemas relativamente simples, porém suficientes para ilustrar as possibilidades em utilizar o sistema em cenários reais. No total, o catálogo da IDE possui 11 metadados e o tesauro utilizado no estudo de caso contém 65 termos. A primeira etapa do teste do protótipo consistiu em gerar base indexada. Antes de gerar a base indexada, é necessário importar os metadados e o tesauro para as bases de recurso e semântica, respectivamente. A Figura 5.6 ilustra os passos executados pelo usuário e as mensagens de retorno do protótipo.



Fonte: Elaborado pelo Autor

Figura 5.6: Geração da Base Indexada.

A Tabela 5.1 apresenta o conteúdo da Base de Recurso após a importação dos metadados, onde a coluna Id corresponde ao identificador do metadado no catálogo da IDE e a coluna Recurso armazena suas palavras-chave. A coluna Item foi adicionada à tabela para facilitar a citação e localização dos itens da tabela no texto.

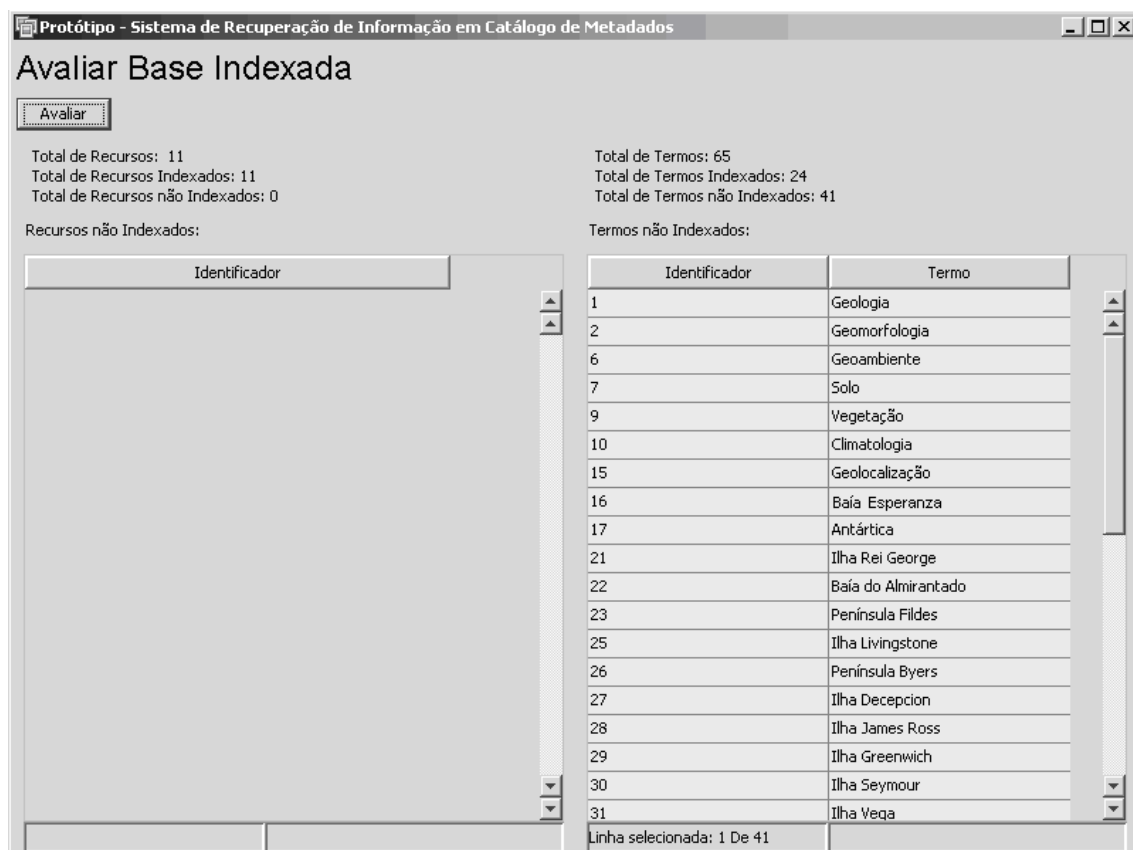
Tabela 5.1: Conteúdo da Base de Recurso

Item	Id	Recurso
1	011329a5-b008-4fd2-9a6c-ad8085e33b44	"Perfil de Solo", "Amostra de Solo", "Península Potter"
2	b6d3c25b-d596-4322-bc96-2c4f27bbc645	"Península Potter", "Flora", "Fauna"
3	7be4fe37-c5e7-4471-bca0-bff11826eaba	"Curva de nível", "Relevo", "Península Potter"
4	fd85430e-5da2-4221-abad-15e18a2cd578	"Península Potter", "Solos Gelados"
5	702114ac-7739-4681-bf1b-738c477195d5	"Potter Peninsula", "Permafrost"
6	6922aa64-04ef-4765-9830-661e30e3b3b5	"Península Potter", "Criossolos"
7	c37ecac0-f7b5-4c67-aa06-bb2c9733cf5c	"Hidrologia", "Drenagem", "Península Potter"
8	b485b990-ceb2-4e77-bf56-a8e5eb2304f5	"Infraestrutura de Dados Espaciais", "Tesauro", "Topic maps", "Recuperação de Informação", "Metadado"
9	b94554e7-56cd-4a93-821b-fc0525131c26	"Spatial Data Infrastructures", "metadata", "ontologies", "Sharing Knowledge", "Knowledge Reuse", "Analysis patterns"
10	6b0c2e82-a5e8-4c11-a507-ec6bc9c42769	"Spatial Data Infrastructures", "metadata", "topic map", "knowledge base", "Thesaurus", "knowledge representation"
11	862fde59-0f60-49a3-a3f8-b3e7ff9dbd16	"Metadado", "Ecossistema Antártico", "Conhecimento geoespacial", "Infraestrutura de dados espaciais"

Fonte: Elaborado pelo Autor

A segunda etapa do teste do protótipo se resume em avaliar a Base Indexada, sendo os resultados apresentados na Figura 5.7. Analisando essa avaliação, pode-se observar que todos os recursos foram indexados. Esse resultado se deve ao processo adotado na criação do metadado e descrito na seção anterior, em que todo metadado

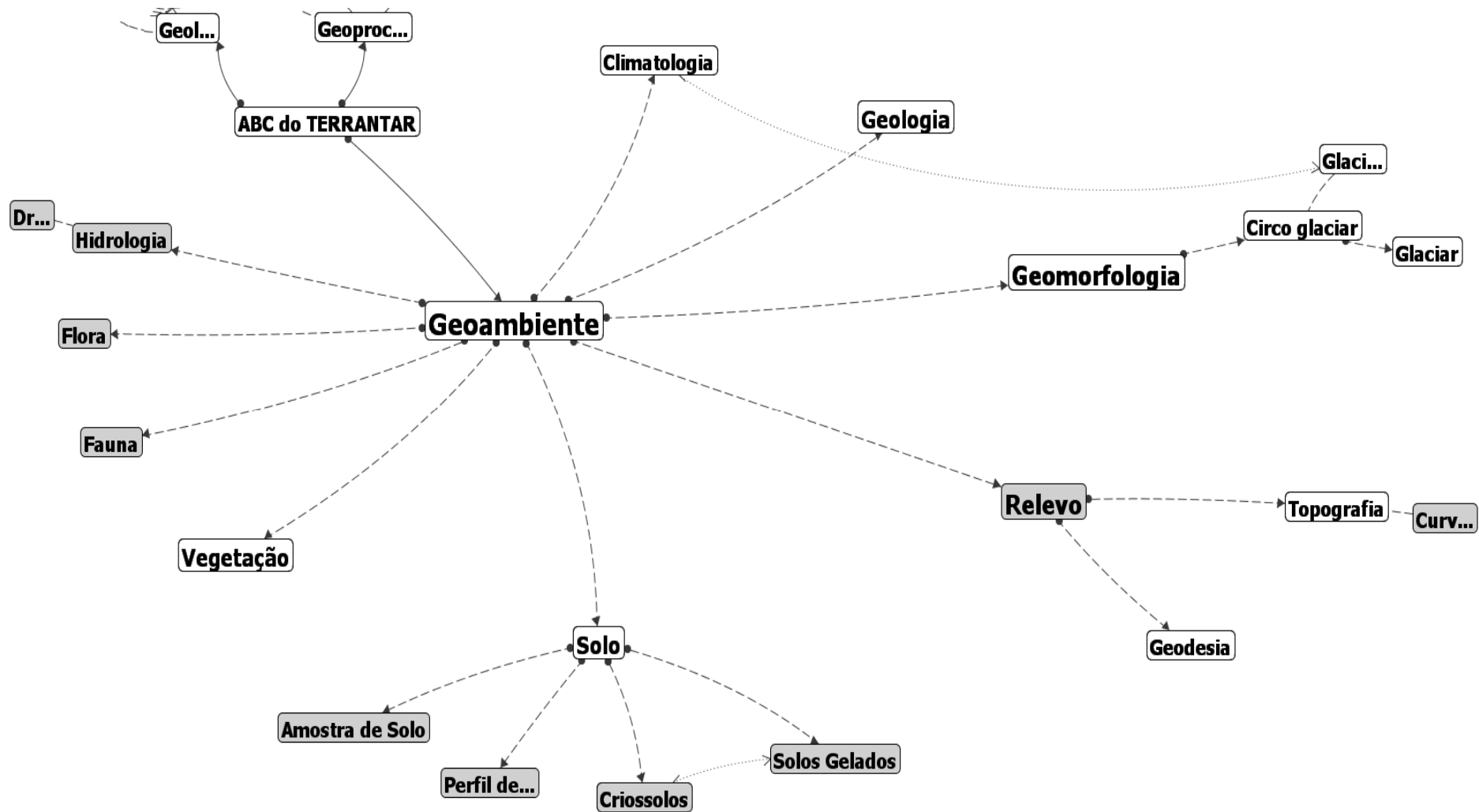
deve consumir termos do tesauro. Outra análise que pode ser feita é em relação aos termos indexados. O fato de existirem termos não indexados já era de se esperar, visto que alguns termos servem de temas/categorias no tesauro, e dificilmente são utilizados na descrição do metadado. Outro fator que contribuiu para a diferença entre os termos existentes e não indexados é que o tesauro já está preparado para receber outros conjuntos de metadados. Uma lista contendo todos os elementos do tesauro importados para a Base Semântica está descrita no Apêndice A.



Fonte: Elaborado pelo Autor

Figura 5.7: Avaliação da Base Indexada.

A terceira e última etapa do teste consistiu em realizar consultas com o protótipo e analisar seus resultados. Por fim, foram submetidas consultas semelhantes à interface padrão da IDE, o que permitiu comparar os dois mecanismos. Uma visão parcial da árvore hiperbólica gerada pelo protótipo é ilustrada pela Figura 5.8.



Fonte: Elaborado pelo Autor

Figura 5.8: Visão parcial da árvore hiperbólica.

Pode-se observar que os termos indexados (ex: “Solos Gelados” e “Amostra de Solos”) têm sua coloração alterada, essa característica, além de evitar consultas desnecessárias, facilita sua identificação pelo usuário. A ligação entre os termos relacionados (ex: “Criossolos” e “Solos Gelados”) tem seu estilo pontilhado, diferenciando-se dos outros relacionamentos. Outra característica a ser destacada é o *multilíngue*, que foi levada em consideração no desenvolvimento da função de consulta. Essa característica não está visível na Figura 5.8, mas pode ser observada a relação de equivalência entre os termos “*Permafrost*” e “Solos Gelados”, no Apêndice A. A seguir, são descritos os testes, as análises dos resultados e as comparações com outro mecanismo de busca.

5.4.1 Teste I: consulta com termos

O primeiro teste simula uma situação em que é informado apenas um termo e nenhum critério de abrangência. A Figura 5.9 ilustra os critérios dessa consulta e seu resultado.

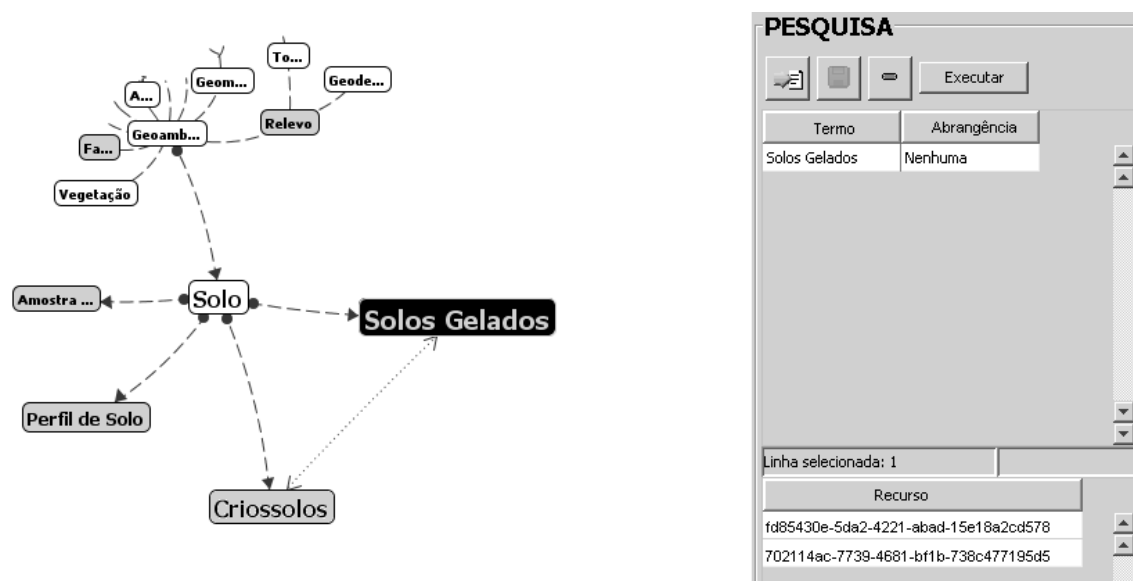


Figura 5.9: Teste I: consulta com termos.

A consulta retorna dois identificadores (Itens: 4 e 5 da Tabela 5.1) de metadados. Analisando o resultado com o conteúdo da Tabela 5.1, é possível verificar que o fator *multilíngue* foi levado em consideração na busca. O termo é passado como parâmetro na opção palavras-chave da interface da IDE, que retorna apenas o identificador (Item: 4). É importante observar o valor da precisão da busca da interface, este critério não foi alterado em nenhum dos testes realizados (ver Figura 5.2).

5.4.2 Teste II: consulta com termos relacionados

No teste II, utilizou-se o mesmo termo anterior e acrescentou-se como critério de abrangência a opção que faz referência a seus termos relacionados. A Figura 5.10 ilustra a consulta e seu resultado.

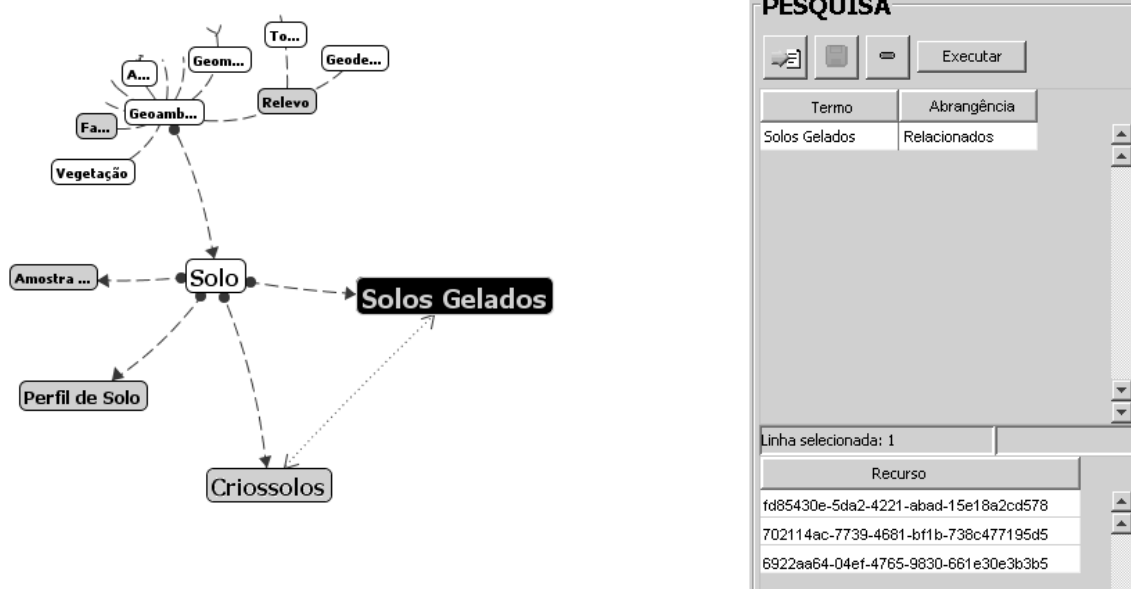


Figura 5.10: Teste II: consulta com termos relacionados

A consulta retorna três identificadores (Itens: 4, 5 e 6 da Tabela 5.1) de metadados. Analisando o resultado com o conteúdo da Tabela 5.1, é possível verificar que o critério da abrangência é utilizado na busca. Como o termo não foi alterado, o resultado exibido na interface da IDE é o mesmo do Teste I.

5.4.3 Teste III: consulta com termos e especializações

O terceiro teste simulou uma situação em que foi informado um termo que não foi indexado, junto com seu critério de abrangência que faz referência a seus termos especializados. A Figura 5.11 ilustra os critérios desta consulta e seu resultado. A consulta retorna quatro identificadores (Itens: 1, 4, 5 e 6 da Tabela 5.1) de metadados. Analisando o resultado, de acordo com o conteúdo da Tabela 5.1, é possível verificar que, além do critério da abrangência, o fator *multilíngue* foi utilizado na busca. Seguindo o procedimento dos testes anteriores, o termo foi informado na opção de palavras-chave da outra interface e retornou apenas um identificador (Item: 1 da Tabela 5.1). Não foi possível analisar a implementação do mecanismo de busca desta interface, mas por analogia com outros testes realizados, acredita-se que o mecanismo reconheceu

a ocorrência do termo “Solo” nas palavras-chave “Perfil de Solo” e “Amostra de Solo”. O mesmo não aconteceu em relação à palavra-chave “Solos Gelados”, caso essa palavra estivesse descrita no singular a mesma seria reconhecida.

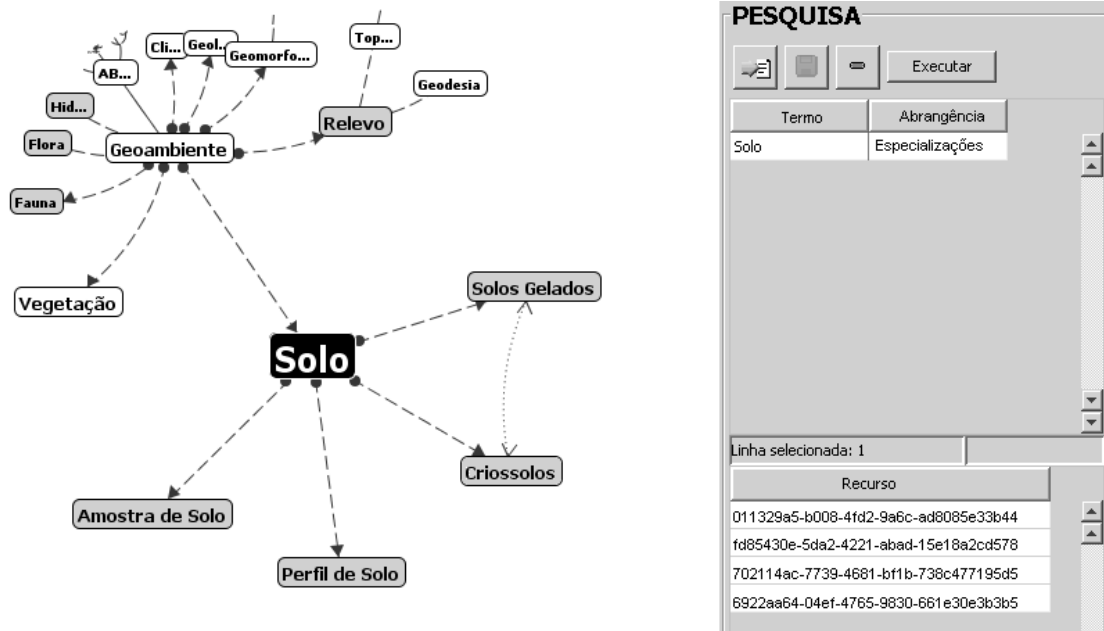


Figura 5.11: Teste III: consulta com termos e especializações.

5.4.4 Comparação e avaliação dos testes do protótipo

Não é objetivo deste trabalho discutir os critérios relacionados à usabilidade das interfaces apresentadas, o que foge ao escopo desta dissertação e exigiria uma pesquisa das técnicas utilizadas na disciplina Interface Homem-Máquina. O que diferencia as duas interfaces e foi levado em consideração no desenvolvimento do protótipo foram os recursos que auxiliam e contribuem para o entendimento do contexto a ser consultado.

A avaliação do protótipo restringe-se a mostrar as possíveis soluções para algumas questões semânticas, que diferenciam essa proposta em relação a um mecanismo de busca semelhante ao adotado pela interface da Figura 5.2. Deve-se observar que esta mesma interface pode consumir um serviço Web que implementa a função de busca utilizada pelo protótipo e apresentar os mesmos resultados dos testes realizados.

6 CONCLUSÕES

Neste capítulo, é feita uma avaliação geral do trabalho desenvolvido. Para isso, os objetivos propostos inicialmente são considerados, assim como os meios utilizados para alcançá-los. Dessa forma, é possível apresentar as principais conclusões e contribuições desta dissertação. Em seguida, são enumeradas algumas sugestões de trabalhos futuros, com a intenção de contribuir com o desenvolvimento desta linha de pesquisa.

A partir dos estudos e pesquisas realizadas, uma arquitetura para sistemas de recuperação de informação em catálogos de metadados foi especificada. A arquitetura proposta descreve um mecanismo de busca que faz uso de tesauro e *topic maps*. Esses artefatos semânticos contribuem para o entendimento do contexto a ser consultado e influenciam o mecanismo de busca. Esta abordagem permitiu minimizar alguns problemas relacionados à semântica no processo de recuperação de informação. Com isso, o objetivo geral, proposto nessa pesquisa, foi atingido.

Os objetivos específicos também foram alcançados. A arquitetura especificada no Capítulo 3 foi utilizada na implementação de um protótipo que permitiu recuperar informação em um catálogo de metadados de uma IDE. A viabilidade do sistema desenvolvido foi verificada por meio do desenvolvimento de um estudo de caso, no contexto de uma IDE para o ecossistema antártico. O protótipo foi testado, utilizando dados reais do catálogo da IDE do estudo de caso, e permitiu demonstrar possíveis soluções para questões semânticas na recuperação de informação.

O protótipo foi desenvolvido com recursos que auxiliam e podem contribuir no entendimento semântico de um contexto. Em termos teóricos, buscou-se simplificar a interação do usuário com a interface de busca. Entretanto, a lógica implementada no mecanismo de busca é a chave para resolver possíveis problemas de semântica na recuperação de informação.

Para auxiliar os produtores de dados geográficos, também usuários da IDE, foi definido um procedimento a ser seguido na catalogação dos metadados. Esse procedimento é composto por atividades e permite uma evolução colaborativa do tesauro utilizado no estudo de caso. A adoção desse procedimento é fundamental para se continuar a obter resultados satisfatórios na utilização do sistema.

O desenvolvimento desta pesquisa apresentou algumas dificuldades. Uma das mais marcantes está relacionada ao caráter interdisciplinar, pois a pesquisa envolveu

competências de várias disciplinas e dedicação de vários colaboradores. Entretanto, essa diversificação de conteúdo e a aplicação em um cenário real enriqueceram a pesquisa. Houve, além disso, dificuldades relacionadas à implementação do protótipo, como o aprendizado da biblioteca Treebolic.

As principais contribuições desta dissertação são:

- Levantamento e comparação de artefatos semânticos para representação do conhecimento, propostos na literatura;
- Especificação e desenvolvimento de uma IDE que atenda às necessidades do projeto TERRANTAR;
- Levantamento das necessidades de uma arquitetura para sistemas de recuperação de informação em catálogos de metadados;
- Especificação de um sistema, com um conjunto de operações, uma estrutura de repositório e um mecanismo de busca que faz uso de tesauro e *topic maps* na recuperação de informação;
- Implementação de um protótipo, com considerações em relação a sua interação semântica com o usuário.

Esta dissertação admite extensões que podem contribuir para o desenvolvimento desta linha de pesquisa. As sugestões de trabalhos futuros estão relacionadas tanto aos aspectos de pesquisa quanto de implementação, algumas das quais estão listadas a seguir:

- Expandir e evoluir as especificações da arquitetura proposta, a fim de que seja oferecido um mecanismo de busca cada vez mais completo para a recuperação de informação em catálogo de metadados;
- Especificar um serviço Web que disponibilize acesso ao mecanismo de busca a outras aplicações clientes;
- Propor um mecanismo que, ao analisar as bases de recurso e semântica, sugira novos candidatos a termos semânticos;
- Especificar e implementar uma nova versão do sistema, levando em consideração a composição do conteúdo da base de recurso, um algoritmo alternativo para geração da base indexada e novos indicadores que auxiliem o usuário na avaliação da base indexada;
- Pesquisar os aspectos temporais relacionados à recuperação de informação em catálogo de metadados.

REFERÊNCIAS BIBLIOGRÁFICAS

- ADITYA, T. **The national atlas as a metaphor for improved use of a national geospatial data infrastructure**. 2007. Tese – Ph.D. Enschede: ITC.
- ADITYA, T.; KRAAK, M.-J. A search interface for an SDI: implementation and evaluation of metadata visualization strategies. **Transactions in GIS**, v. 11, n. 3, p. 413-435, 2007.
- AHMED, K. **Topic map design patterns for information architecture**. XML 2003. Disponível em: <<http://www.techquila.com/tmsinia.html>>. Acesso em: dezembro de 2003.
- ATHANASIS, N.; KALABOKIDIS, K.; VAITIS, M.; SOULAKELLIS, N. **Towards a semantics-based approach in the development of geographic portals**. [S.l.]: Computers and Geosciences, 2008. In press (doi: 10.1016/j.cageo.2008.01.014).
- BERNERS-LEE, T.; HENDLER, J.; AND LASSILA, O. The semantic Web. **Scientific American**, v. 284, n.5, p. 34-43, Maio 2001.
- BOOCH, G.; RUMBAUGH, J.; JACOBSON, I. **Unified modeling language user guide**. 2. ed. [S.l. : s.n.], 2005. ISBN: 0321267974.
- BURKE, M. A. **Organization of multimedia resources**: principle and practice of information retrieval. Aldershot: Gower, 1999.
- BURROUGH, P. A.; MCDONNELL, R. A. **Principles of geographical information systems**. Oxford: Oxford University Press, 1998. 333 p.
- BUZAN, T. **The mind map book**: how to use radiant thinking to maximize your brain's untapped potential. [S.l.]: Plume Books, 1996.
- CAMPOS, M. L. A. et al. Estudo comparativo de softwares de construção de tesauros. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 1, p. 68-81, 2006.
- CASANOVA, M. et al. Integração e interoperabilidade entre fontes de dados geográficos. In: **Banco de dados geográficos**. [S.l.]: MundoGeo, 2005. p. 305-340. Disponível em: <<http://www.dpi.inpe.br/gilberto/livro/bdados>>.
- CRAGLIA, M.; GOODCHILD, M. F. "Next-Generation Digital Earth." **International Journal of Spatial Data Infrastructures Research**, v. 3, p. 146-167, 2008.

- CROMPVOETS, J.; BREGT, A.; RAJABIFARD, A.; WILLIAMSON, I. Assessing the worldwide developments of national spatial data clearinghouses. **International Journal of Geographic Information Science**, v. 18, n. 7, p. 665-689, 2004.
- DALTIO, J.; MEDEIROS, C. M. B. Aondê: **An ontology Web service for interoperability across biodiversity applications**. *Information Systems*, v. 33, p. 724-753, 2008.
- DAVENPORT, T. H.; PRUSAK, L. **Conhecimento empresarial: como as organizações gerenciam seu capital intelectual**. Rio de Janeiro: Campus, 1998.
- DAVIS JR., C. A.; ALVES, L. L. Local spatial data infrastructures based on a service-oriented architecture. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS. **Proceedings...** [S.l. : s.n.], 2005. p. 30-45.
- DAVIS JR., C. A.; FONSECA, F. Understanding global change: the role of geographic information science in the integration of people and nature. In: NYERGES, T.; COUCLELIS, H.; McMASTER, R. B. **Handbook of GIS and society SAGE**. [S.l. : s.n.], 2009.
- DAVIS JR. et al. Infraestruturas de dados espaciais na integração entre ciência e comunidades para promover a sustentabilidade ambiental. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS, 1., 2009, Bento Gonçalves. **Anais...** Porto Alegre: Sociedade Brasileira de Computação, 2009.
- DODEBEI, Vera Lúcia Doyle. **Tesouro: linguagem de representação da memória documentária**. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002.
- EGENHOFER, M. Toward the semantic geospatial Web. In: ACM INTERNATIONAL SYMPOSIUM ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 10., 2002, Virgínia. **Proceedings...** McLean, Virginia: ACM Press, 2002.
- FENSEL, D. et al. OIL: an ontology infrastructure for the semantica web. **IEEE Intelligent Systems**, p. 38-45, Mar./Apr. 2001.
- FERNEDA, E. **Recuperação da informação: análise sobre a contribuição da ciência da computação para a ciência da informação**. 2003. 147 f. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Arte da Universidade de São Paulo, São Paulo, 2003.

- FGDC. Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata. **Federal Geographic Data Committee**. Washington: D.C. 1998. 78p.
- GARSHOL, L. M. Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. **Journal of Information Science**, v. 30, n. 4, p. 378-391, 2004.
- GRAND, B. L.; SOTO, M. Visualisation of the semantic Web: topic maps visualisation. In: **International Conference on Information Visualisation (IV'02)**, 6., p. 344-349, Jul. 2002.
- GUARINO, N. Formal ontology and information systems. In: **INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY IN INFORMATION SYSTEMS**, 1., 1998, Trento. **Proceedings...**Trento, Italy, Jun. 1998.
- GRUBER, T. R. Towards principles for the design of ontologies used for knowledge sharing. **International Journal of Human-Computer Studies**, v. 43, n. 5-6, p. 907-928, 1995.
- HOCHMAIR, H. Ontology matching for spatial data retrieval from Internet portals. In: **GEOSPATIAL SEMANTICS – Lecture Notes in Computer Science 3799**, 2005, Mexico City. **Proceedings...** Mexico City, Mexico, 2005. p. 166-182.
- ISO 19115. **Geographic information – Metadata**. Geneva: International Organization for Standardization (ISO), 2003.
- ISO 13250. **Topic maps**. 2. ed. [S.l.]: International Organization for Standardization (ISO), 2002.
- ISO 2788. **Guidelines for the establishment and development of monolingual thesauri**. [S.l.]: International Organization for Standardization, 1986.
- ISO 5964. **Guidelines for the establishment and development of multilingual thesauri**. [S.l.]: International Organization for Standardization, 1985.
- JANSEN, B. J.; SPINK, A.; SARACEVIC, T. Real life, real users, and real needs: a study and analysis of user queries on the web. **Information Processing and Management**, v. 36, n. 2, p. 207-227, 2000.
- JASPER, R.; USCHOLD, M. A. Framework for understanding and classifying ontology applications. In: **INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL**

- INTELLIGENCE (IJCAI-99), 1999, Stockholm, Sweden. **Proceedings...** Stockholm, Sweden, July 31 - August 6, 1999.
- KUHLTHAU, C. Inside the search process: information seeking from the user's perspective. **Journal of the American Society for Information Science**, v. 42, n. 5, p. 361-371, 1990.
- LAMPING, J.; RAO, R.; PIROLI, P. A. Focus+context technique based on hyperbolic geometry for visualizing large hierarchies, human factors in computing systems, CHI '95 Conference. **Proceedings...** [S.l.]: ACM Press, 1995. p. 401-408.
- MAEDCHE, STAAB; MAEDCHE, A.; STAAB, S. **A ontology learning for the semantic Web**. [S.l.]: IEEE Intelligent Systems March/April 2001. p. 72-79.
- MAGUIRE, D. J.; LONGLEY, P. A. The emergence of geoportals and their role in spatial data infrastructures. **Computers, Environment and Urban Systems**, n. 29, p. 3-14, 2005.
- MASSER, I. All shapes and sizes: the first generation of national spatial data infrastructures. **International Journal of Geographic Information Science**, v.13, n. 1, p. 67-84, 1999.
- MIGUEL, J. L. **Contributions to the problem of knowledge management in Spatial Data Infrastructures**. [S.l. : s.n.t.].
- MMA – MINISTÉRIO DO MEIO AMBIENTE. **O Brasil e o meio ambiente antártico**. Brasília: MMA, 2007.
- MOREIRA, A. **Tesouros e ontologias**; estudo de definições presentes na literatura das áreas das Ciências da Computação e da Informação, utilizando-se o Método Analítico-Sintético. 2003. 150 f. Dissertação (Mestrado de Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2003.
- MOREIRA, A.; ALVARENGA, L.; OLIVEIRA, A. P. O nível do conhecimento e os instrumentos de representação: tesouros e ontologias. **Datagrama zero: Revista de Ciência da Informação**, v. 5, n. 6, dez. 2004. Disponível em: <www.dgzero.org/dez04/Art_01.htm>.
- NEBERT, D. D. (Ed.). **Developing spatial data infrastructures: the SDI cookbook**, Version 2.0. 2004 (GSDI-Technical Working Group) (available at: www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf).

- NOGUERAS-ISO, J.; ZARAZAGA-SORIA, F. J.; MURO-MEDRANO, P. R. **Geographic information metadata for spatial data infrastructures: resources, interoperability and information retrieval**. [S.l.]: Springer, 2005.
- PEPPER, S. **The TAO of topic maps** – Finding the way in the age of infoglut. Ontopia. 2000. Disponible in: <<http://www.ontopia.net/topicmaps/materials/tao.html>>.
- PONTES, F. V. et al. SBICafé: uma ferramenta de gestão do conhecimento para a comunidade de pesquisa do café. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 2., 2001, Vitória. **Anais...** Vitória, 2001.
- RATH, H. H. **The topic maps handbook**. Gütersloh, Germany: Empolis, 2003.
- RESENDE, M.; CURI, N.; REZENDE, S. B. D.; CORRÊA, G. F. **Pedologia: base para distinção de ambientes**. Viçosa, MG: NEPUT, 1995. 304 p.
- SIGEL, A. **Towards knowledge organization with topic maps**. [S.l.]: XML Europe, 2000.
- SAVOY, J.; PICARD, J. Retrieval effectiveness on the Web. In: **Information Processing and Management**, v. 37, n. 4, p. 543-569, 1998.
- SOWA, J. F. **Building, sharing and merging ontologies**. [S.l. : s.n.], 2001. 21 p. Disponível em: <<http://www.jfsowa.com/ontology/ontoschar.htm>>.
- VENANCIO, L. R.; FILETO, Renato; MEDEIROS, Claudia Bauzer. **Aplicando ontologias de objetos geográficos para facilitar navegação em GIS**. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA, 2003, Campos do Jordão. GEOINFO, 2003.
- WAZLAZWICK, R. S. **Análise de projeto de sistemas de informação orientados a objetos**. Rio de Janeiro: Elsevier, 2004.
- WELTY, C.; GUARINO, N. Supporting ontological analysis of taxonomic relationships. **Data and Knowledge Engineering**, v. 39, p. 51-74, 2001.

APÊNDICE A

A.1 Elementos do Tesauro – ABC do TERRANTAR

Os tipos de atributos são identificados pela legenda abaixo:

- EQ: termo equivalente;
- TE: termo específico;
- TG: termo genérico;
- TR: termo relacionado;
- USE: termo usado por.

Termos	Atributos
Amostra de Solo	TG: Solo
Antártica	TG: Geolocalização TE: Baía do Almirantado TE: Baía Esperanza TE: Ilha Decepcion TE: Ilha Greenwich TE: Ilha James Ross TE: Ilha Livingstone TE: Ilha Rei George TE: Ilha Seymour TE: Ilha Vega TE: Península Barton TE: Península Byers TE: Península Fildes TE: Península Potter
Baía do Almirantado	TG: Antártica
Baía Esperanza	TG: Antártica
Circo glacial	TG: Geomorfologia TE: Glaciação TE: Glaciar EQ: Cirque glacier
Climatologia	TR: Glaciação TG: Geoambiente
Conhecimento geoespacial	TG: Geoprocessamento TE: Ontologia TE: Tesauro TE: <i>Topic maps</i>
Criossolos	TR: Solos Gelados TG: Solo
Curva de nível	TG: Topografia EQ: Contour line EQ: Isohypse
Drenagem	TG: Hidrologia EQ: Drainage
Fauna	TG: Geoambiente
Flora	TG: Geoambiente
Geleira	USE: Glaciar

Geoambiente	TE: Climatologia TE: Fauna TE: Flora TE: Geologia TE: Geomorfologia TE: Hidrologia TE: Relevô TE: Solo TE: Vegetação
Geodesia	TG: Relevô EQ: Geodesy
Geolocalização	TE: Antártica
Geologia	TG: Geoambiente
Geomorfologia	TG: Geoambiente TE: Circo glaciár
Geoprocessamento	TE: Conhecimento geoespacial TE: Infraestrutura de Dados Espaciais TE: Metadado
Glaciação	TR: Climatologia TG: Circo glaciár EQ: Glaciation
Glaciár	TG: Circo glaciár UP: Geleira EQ: Glacier
Hidrologia	TG: Geoambiente TE: Drenagem EQ: Hydrology
Ilha Decepcion	TG: Antártica
Ilha Greenwich	TG: Antártica
Ilha James Ross	TG: Antártica
Ilha Livingstone	TG: Antártica
Ilha Rei George	TG: Antártica
Ilha Seymour	TG: Antártica
Ilha Vega	TG: Antártica
Infraestrutura de Dados Espaciais	TG: Geoprocessamento EQ: Spatial Data Infrastructures
Metadado	TG: Geoprocessamento EQ: Metadata
Ontologia	TG: Conhecimento geoespacial EQ: Ontologies
Península Barton	TG: Antártica
Península Byers	TG: Antártica
Península Fildes	TG: Antártica
Península Potter	TG: Antártica EQ: Potter Peninsula
Perfil de Solo	TG: Solo
Relevô	TG: Geoambiente TE: Geodesia TE: Topografia

	EQ: Relief
Solo	TG: Geoambiente TE: Amostra de Solo TE: Criossolos TE: Perfil de Solo TE: Solos Gelados EQ: Soil
Solos Gelados	TR: Criossolos TG: Solo EQ: Permafrost
Tesouro	TG: Conhecimento geoespacial EQ: Thesaurus
<i>Topic maps</i>	TG: Conhecimento geoespacial
Topografia	TG: Relevo TE: Curva de nível EQ: Topography
Vegetação	TG: Geoambiente