LIGIANE ALVES DE SOUZA

LOCUS: Um Sistema de Localização Geográfica Através de Referências Espaciais Indiretas

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Belo Horizonte

2005

Agradecimentos

A Deus, que tornou tudo possível.
A Iglesias, meu amor.
A minha mãe, meu pai e meus irmãos, que trago sempre no coração.
Ao prof. Clodoveu, que com inteligência e criatividade me indicou o caminho certo que deveria seguir.
Ao prof. Laender, pelo apoio em todos os momentos.
À Karla, essa pessoa de alma tão bonita, que esteve ao meu lado em todos os passos que dei ao longo desta caminhada.
Aos amigos Fabiano, Wagner, Luciano e Marcos, pela alegria da convivência.
Obrigado, por tudo!

Resumo

Uma *referência espacial indireta* é qualquer elemento textual capaz de identificar um lugar sem fazer uso explícito de suas coordenadas geográficas. Nomes de lugares, endereços, códigos postais e números de telefone são alguns exemplos de referências espaciais indiretas. Um *gazetteer* é um dicionário de nomes de lugares. Sua utilidade é informar a localização geográfica de referências espaciais indiretas. Nos últimos anos, alguns *gazetteers* digitais têm surgido com o objetivo principal de apoiar atividades nas áreas de recuperação de informação geográfica e georreferenciamento de dados. Esta dissertação apresenta o Locus, um sistema de localização através de referências espaciais indiretas. O Locus é fundamentado em um *gazetteer* digital, cuja modelagem seguiu os conceitos definidos em uma ontologia, a *ontologia de lugar*. Essa metodologia de desenvolvimento resultou na construção de um *gazetteer* mais elaborado e poderoso que os disponíveis atualmente. Na formação do conteúdo inicial do Locus, foram utilizadas fontes de dados urbanos, regionais e nacionais, obtidos em organizações de natureza bastante diversa, como os Correios, a ANATEL, o IBGE e a Prefeitura de Belo Horizonte. A esses dados foram acrescentados 29.139 nomes de lugares obtidos diretamente a partir da Internet, usando um mecanismo de coleta baseado em exemplos. Esta dissertação inclui uma avaliação do potencial de uso do Locus em uma aplicação de recuperação de informação geográfica, e propõe aplicações do localizador em diferentes áreas, tanto por meio de uma interface interativa quanto através de *Web services*.

Abstract

An *indirect spatial reference* is any textual element used to identify places, without the use of spatial or geographic coordinates. Place names, urban addresses, postal area codes, and telephone numbers are examples of indirect spatial references. A gazetteer is a dictionary of place names. Its purpose is to supply the actual spatial location of indirect geographic references. In the recent past, some digital gazetteers have been developed to support geographic information retrieval initiatives and to enhance georeferencing applications. This dissertation presents Locus, a spatial locator system based on indirect geographic references. Locus includes a digital gazetteer, which was modeled in accordance to an ontology, the *ontology of places*. With this, the Locus gazetteer has become more elaborated and powerful than other currently available gazetteers. In the creation of Locus' initial contents, data sources at the urban, regional, and national levels were employed, including the Brazilian postal company, the Brazilian telecommunications agency, and the local government for the city of Belo Horizonte. Furthermore, 29.139 place names were added, after being collected semi-automatically from Web sources using a data extraction by example mechanism. This dissertation also includes an evaluation of the potential use of Locus in a geographic information retrieval context, and proposes further applications in various fields, using either an interactive interface or Web services.

Sumário

1 Introdução 1
2 Trabalhos Relacionados 4
2.1 Recuperação de Informação Geográfica 4
2.2 Georreferenciamento 12
2.3 Gazetteers Digitais 14
3 Projeto do Locus 19
3.1 Ontologia de Lugar 19
3.2 Esquema Conceitual OMT-G do Gazetteer 23
3.3 Modelagem do Locus 26
4 Implementação e Avaliação do Locus 30
4.1 Opções de implementação 30
4.2 Carga de Dados 31
4.3 Implementação do Módulo de Busca 38
4.4 Implementação do Módulo de Ranking de Lugares 40
4.5 Interface 42
4.6 Avaliação 48
5 Conclusões e Trabalhos Futuros 51
6 Referências bibliográficas 54
Apêndice – Notação OMT–G 59

Índice de Figuras

Figura 2.1 – Arquitetura do SPIRIT. Fonte: [14]. 8
Figura 2.2 – Esquema OMT-G da ontologia de lugar do SPIRIT. Fonte: [16] 8
Figura 2.3 – Identificação do contexto geográfico pela Tumba. Fonte: [47] 9
Figura 2.4 – Gramática para consultas espaciais. 11
Figura 3.1 – Abordagem dirigida por ontologia para a modelagem de SIGs. Fonte: [15] 21
Figura 3.2 – Ontologia de lugar. Fonte: [7] 22
Figura 3.3 – Esquema conceitual OMT-G do *gazetteer* do Locus. 23
Figura 3.4 – Esquema lógico do *gazetteer*. 25
Figura 3.5 – Diagrama de casos de uso. 26
Figura 3.6 – Diagrama de seqüência do caso de uso Consulta simples. 27
Figura 3.7 – Diagrama de seqüência do caso de uso Consulta avançada. 28
Figura 3.8 – Arquitetura do sistema. 29
Figura 4.1 – Interface gráfica da ASByE. 34
Figura 4.2 – Interface gráfica da DEByE. 35
Figura 4.3 – Algoritmo para determinação da similaridade entre duas referências. 36
Figura 4.4 – Processo de alimentação do *gazetteer* pela Web. 36
Figura 4.5 – Tela de consulta simples. 42
Figura 4.6 – Tela de seleção da consulta simples. 43
Figura 4.7 – Tela de resultado da consulta simples. 43
Figura 4.8 – Tela de consulta avançada. 44
Figura 4.9 – Tela de seleção do ponto de referência na consulta avançada. 44
Figura 4.10 – Tela de seleção do ponto de interesse na consulta avançada. 45
Figura 4.11 – Tela de consulta de endereço. 45
Figura 4.12 – Tela de resultado da consulta de endereço. 46
Figura 4.13 – Tela de consulta avançada com endereço. 47
Figura 4.14 – Tela de resultado da consulta avançada com endereço. 47
Figura 4.15 – Questionário de avaliação do experimento. 49
Figura 4.16 – Avaliação da relevância das respostas do Google. 49
Figura 4.17 – Sucesso do Locus em encontrar os lugares indicados nas consultas. 50

Índice de Tabelas

Tabela 2.1 – Alguns *gazetteers* disponíveis na Web. 16
Tabela 4.1 – Carga inicial de dados no *gazetteer* do Locus. 32
Tabela 4.2 – Resultados da coleta de páginas e da extração de referências. 37
Tabela 4.3 – Classificação do nível de popularidade de uma referência. 41
Tabela 4.4 – Mapeamento entre o nível de popularidade de uma referência e o *ranking* de lugares. 41
Tabela 4.5 – Simulação de valores para o índice final usado na ordenação dos resultados. 42

1 Introdução

Nos últimos anos, a pesquisa na área de recuperação de informação geográfica tem alcançado um destaque considerável nos meios acadêmico e comercial. O motivo principal desse interesse é a constatação de que os métodos atuais de recuperação de informação, baseados principalmente em busca por palavras-chave e análise de *links*, não são suficientes para atender à necessidade dos usuários de localizar informações com contexto espacial. Iniciativas como o projeto SPIRIT (http://www.geo-spirit.org) e o Google Local (http://local.google.com), demonstram uma demanda crescente por serviços de busca que considerem a semântica geográfica dos documentos.

Por outro lado, o georreferenciamento automático de registros é um recurso bastante útil em sistemas de informação geográficos (SIG). Muitos bancos de dados não são georreferenciados, mas possuem em seus registros atributos que tornam possível um georreferenciamento automático, como endereços semi-estruturados.

Várias evidências podem ser consideradas para definir o contexto geográfico de um documento ou de um registro, como endereços, códigos postais, números de telefone e nomes de pontos de referência. Uma vez identificada uma evidência geográfica, também chamada de *referência espacial indireta* [21], e solucionadas ambigüidades semânticas que possam surgir, é possível associar uma posição (mesmo que aproximada) na superfície terrestre ao documento ou registro em questão e indexá-lo com base em critérios geográficos.

Para a identificação de referências espaciais indiretas, é muito comum a utilização de *gazetteers*. Um *gazetteer* é um catálogo de nomes de lugares (um dicionário toponímico) capaz de fornecer um vocabulário de termos geográficos, acompanhados de suas respectivas localizações [22]. Tradicionalmente, um *gazetteer* refere-se a localizações usando a identificação de um mapa, juntamente com uma referência aproximada, como o número de uma quadrícula. Mais recentemente, a disponibilidade de dados geográficos em meio digital tem permitido a criação de *gazetteers* em que a localização é bem mais precisa, e com recursos de navegação entre objetos geográficos relacionados entre si [1]. Com isso, *gazetteers* podem assumir um papel importante na arquitetura de bibliotecas digitais georreferenciadas [22]. Bibliotecas dessa natureza

exploram a associação que pode existir entre seus objetos e algum lugar na superfície terrestre, criando novas possibilidades de busca, navegação, visualização e acesso aos objetos.

O objetivo deste trabalho consistiu na concepção e implementação do Locus [50], um sistema de localização através de referências espaciais indiretas, no qual ontologias de lugar são apoiadas por um *gazetteer*, gerando recursos de busca e navegação inovadores. Algumas contribuições importantes do Locus são: a capacidade de processamento de consultas espaciais, o armazenamento de referências intra-urbanas, a possibilidade de alimentação do gazetteer com dados extraídos da Web e a busca a nomes de lugares admitindo erro na cadeia de caracteres de consulta. O Locus encontra-se disponível na Web, no endereço http://www.lbd.dcc.ufmg.br:8080/locus.

A motivação principal e mais imediata para a implementação do sistema veio do projeto em desenvolvimento no Laboratório de Bancos de Dados (LBD) do Departamento de Ciência da Computação da UFMG, que estuda o reconhecimento de evidências geográficas em páginas Web e de como essa informação pode ser empregada na indexação e busca dos documentos [8, 28, 50].

O Locus, porém, tem potencial para ser utilizado como componente de muitas outras aplicações onde o georreferenciamento de documentos e entidades seja uma necessidade. Isso inclui projetos de geoprocessamento mantidos por empresas privadas ou instituições públicas. Um exemplo é o projeto SAUDAVEL (Sistema de Apoio Unificado para Detecção e Acompanhamento em Vigilância Epidemiológica) [46], cujo objetivo principal é o desenvolvimento de ferramentas para auxiliar no controle de endemias e também da criminalidade urbana. Uma das necessidades do projeto é justamente o georreferenciamento de casos de doenças e de ocorrências policiais com base no endereço.

Esta dissertação está organizada como se segue. No Capítulo 2 é apresentado um estudo da literatura abordando os temas recuperação de informação geográfica, georreferenciamento e gazetteers digitais. O Capítulo 3, aborda as atividades desenvolvidas durante a idealização e modelagem do sistema, que foi fortemente baseado em uma ontologia de lugar previamente especificada [7]. No Capítulo 4 são apresentados detalhes relacionados à implementação do sistema, como seus principais módulos e os procedimentos executados para a carga dos dados, assim como também os resultados de uma avaliação feita com o objetivo de estimar a eficácia do sistema.

Finalmente, o Capítulo 5 apresenta as conclusões da dissertação e propõe novos desdobramentos para o trabalho que foi desenvolvido.

2 Trabalhos Relacionados

O objetivo deste capítulo é apresentar um estudo da literatura existente acerca de *gazetteers* digitais. Antes de abordar esse tema, dois tópicos relacionados são discutidos: recuperação de informação geográfica e georreferenciamento.

Recuperação de informação geográfica é um ramo de pesquisa relativamente recente que estuda a exploração do contexto espacial presente em documentos textuais. *Georreferenciamento*, bastante comum nos sistemas de informação geográficos, é o procedimento de atribuição de coordenadas geográficas a uma entidade ou evento espacial. A construção de *gazetteers* digitais, abordada no tópico final do capítulo, pode auxiliar sobremaneira na execução dessas atividades, que podem ser incluídas entre suas principais áreas de aplicação.

2.1 Recuperação de Informação Geográfica

Diferentes dimensões podem ser utilizadas para a organização, indexação, busca e navegação de páginas Web [33]. Um usuário pode, por exemplo, estar interessado apenas em documentos escritos em japonês ou somente naqueles modificados após certa data. Dessas dimensões de pesquisa, uma das mais importantes é a que envolve a recuperação de informação baseada em critérios geográficos. Somos todos grandes "consumidores" de informação geográfica: a todo o momento há viajantes precisando de informações sobre seus destinos, empresas procurando áreas para expandir seus negócios, governos estudando o impacto de medidas sociais sobre suas regiões. *Recuperação de informação geográfica* é a área que estuda soluções para suprir essa necessidade de informação espacializada.

Recuperação de informação geográfica é uma área de pesquisa aplicada que combina as tecnologias de sistemas de gerenciamento de bancos de dados (SGBD), interface humano-computador (IHC), sistemas de informação geográficos (SIG) e recuperação de informação (RI), e diz respeito à indexação, busca, recuperação e navegação em fontes de informação georreferenciada, e também ao projeto de sistemas para executar essas tarefas de forma eficaz e eficiente [31].

As fontes de informação disponíveis na Internet podem ser exploradas segundo dois contextos geográficos [33]. O primeiro é chamado de contexto baseado em entidade, em que se considera a localização geográfica do usuário e do servidor na rede. Bancos de dados de registros de domínios, como o Whois (www.whois.net), são capazes de fornecer o endereço físico de servidores na Internet. Partes dos endereços de servidores DNS (*Domain Name Server*) também podem ajudar na determinação da localização geográfica. O Instituto de Pesquisa de Stanford propôs recentemente um domínio de alto nível (.geo) para o georreferenciamento de servidores na Internet, que se encontra em análise no ICANN (*The Internet Corporation for Assigned Names and Numbers*) [24] . Segundo a proposta, a superfície terrestre seria dividida em uma grade de células, e os nomes do domínio seriam referências às essas células. Domínios já existentes poderiam se registrar em um domínio .geo, tornando-se entidades georreferenciadas. Além disso, uma RFC (*Request for Comments*) propondo a inclusão de um registro geográfico no protocolo DNS encontra-se atualmente em estudo na *Internet Engineering Task Force* (IETF) [13].

A exploração do contexto geográfico baseado em entidades torna possível o fornecimento de serviços direcionados para a região onde o usuário se encontra. O *marketing* de produtos, por exemplo, pode ser feito de forma muito mais eficiente e lucrativa se esse recurso for empregado.

O segundo contexto é ainda mais poderoso, e se baseia no conteúdo dos documentos disponíveis na rede. Muitos recursos na Web possuem algum tipo de informação espacial. A página institucional de uma empresa, por exemplo, pode conter seu endereço postal. Uma reportagem divulgada no sítio de um jornal pode fazer referência à cidade onde o fato noticiado ocorreu. Nomes de lugares, endereços, códigos postais e números de telefone são elementos muito comuns em páginas Web e que podem ser utilizados na determinação do contexto geográfico dos documentos.

Muitos usuários procuram por esses recursos quando utilizam máquinas de busca. Em [45], é apresentada uma análise do *log* das consultas submetidas à máquina de busca Excite (http://www.excite.com) durante o ano de 2001. A quantidade e a natureza das consultas que continham algum termo geográfico foram levantadas no estudo. Os termos geográficos considerados foram nomes de lugares, códigos postais, adjetivos que se referem a lugares (“americano”, “internacional”), tipos de lugar e direções (“norte”, “sul”). O estudo apontou que uma parcela significativa das consultas (18,6%) continha termos dessa natureza. Um percentual aproximado a este (14,1%) foi encontrado em

uma avaliação do *log* de consultas da máquina de busca brasileira TodoBR[1] (vide Capítulo 4.6).

Infelizmente, o uso isolado das técnicas de indexação de documentos empregadas atualmente pelas principais máquinas de busca disponíveis na Web tem se mostrado insuficiente em saciar a necessidade de informação geográfica de seus usuários. Os métodos de recuperação de informação atuais são limitados à busca por palavras-chave e ignoram o conteúdo semântico dos documentos [11]. Muitas páginas na Web possuem algum tipo de informação que pode ser empregada para contextualizar a página geograficamente, como, por exemplo, um endereço ou um número de telefone. Essa informação, porém, não é aproveitada na elaboração dos índices de busca. A conseqüência disso é que o usuário freqüentemente não consegue encontrar a informação que está procurando, ou tem que filtrar manualmente em um conjunto muito mais amplo de respostas, muitas das quais não o satisfazem geograficamente. Métodos mais adequados para a recuperação de informação geográfica precisam ser desenvolvidos.

A comunidade envolvida com a área de recuperação de informação tem demonstrado um interesse crescente no tema. Em paralelo à conferência ACM SIGIR 2004, ocorreu o primeiro Workshop on Geographic Information Retrieval. O relatório final do workshop relaciona alguns problemas considerados grandes desafios na área e que ainda merecem muita pesquisa [43]. São eles:

- a extração de termos geográficos em textos não estruturados;
- a identificação e remoção de ambigüidades no processo de extração;
- o estudo de técnicas para o armazenamento eficiente da localização dos objetos e de seus relacionamentos;
- o desenvolvimento de novos algoritmos para que as máquinas de busca tomem proveito da informação geográfica na elaboração de seus índices;
- a combinação da informação geográfica e textual no cálculo da relevância dos documentos; e
- o desenvolvimento de técnicas para a interação do usuário com os resultados das consultas geográficas.

Dentre as iniciativas recentes de se solucionar alguns desses problemas, destaca-se o projeto SPIRIT - *Spatially-aware Information Retrieval on the Internet* (http://geo-

[1] TodoBR: http://www.todobr.com.br

spirit.org) [27]. Patrocinado por um órgão de pesquisa da Comunidade Européia, o SPIRIT encontra-se em fase de desenvolvimento e tem por objetivo atender a necessidade de dois grupos de usuários. O primeiro é o dos que desejam localizar informações sobre um tópico de interesse ou fenômeno que ocorre ou está associado a algum lugar no espaço. Aí estão incluídas, por exemplo, pessoas que procuram por serviços como lojas, cinemas ou restaurantes. O segundo grupo compreende os usuários de sistemas de informação geográficos que estão interessados em algum tipo de informação geoespacial em meio digital, como mapas, imagens ou modelos de terrenos.

A arquitetura do SPIRIT é apresentada na Figura 2.1. Na fase de pré-processamento, as páginas são coletadas e os metadados dos documentos são extraídos. Essa extração compreende a identificação e caracterização dos termos geográficos. Na seqüência, um índice híbrido (textual e espacial) é calculado sobre os documentos. O índice textual é construído baseado na técnica já consagrada de listas invertidas. Diferentes métodos de acesso a dados espaciais, como árvores-R [20] e *quadtrees* [44], estão sendo considerados para o índice espacial.

Já na fase de processamento da consulta, esta é analisada e seus termos geográficos identificados. Esses termos passam por um procedimento de expansão, onde nomes alternativos e lugares geograficamente relacionados são adicionados. A busca é então executada e um *ranking* de relevância geográfica dos documentos é calculado. Esse *ranking* considera principalmente a semelhança entre a geometria do lugar identificado na consulta com a do lugar ao qual o documento se refere.

O componente mais importante da arquitetura do SPIRIT é a sua ontologia de lugar [16]. O propósito da ontologia é formalizar o conhecimento geográfico que é necessário em várias fases operacionais da máquina de busca: extração de metadados, indexação, interpretação e formulação da consulta, e elaboração do *ranking* de relevância. A ontologia proposta pelo SPIRIT (Figura 2.2) faz uma classificação hierárquica dos lugares pelo tipo, tal qual um tesauro. Todo lugar possui um nome padrão e pode possuir também nomes alternativos. Uma ou mais representações geométricas podem estar associadas a um lugar. Essas representações dependem da natureza do lugar e podem assumir a forma de pontos, linhas ou polígonos. As relações espaciais exploradas pelo SPIRIT são apenas “adjacente a”, “sobrepõe”, “contém” e “parte de”.

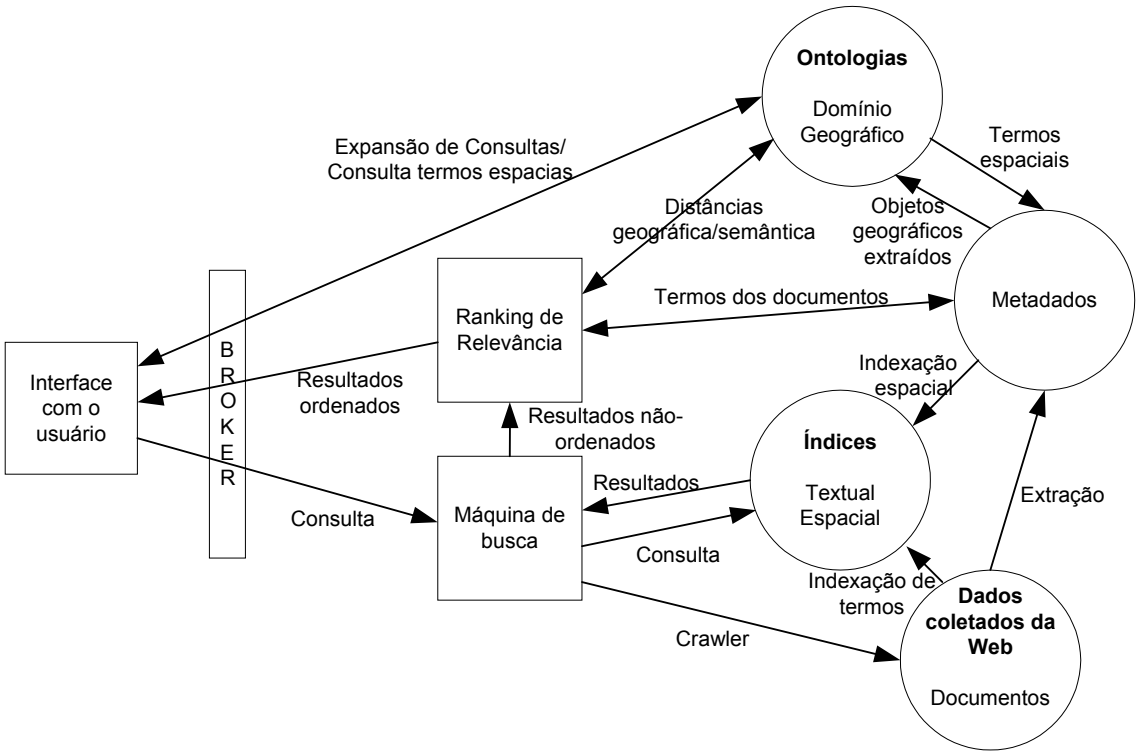


Figura 2.1 – Arquitetura do SPIRIT. Fonte: [14].

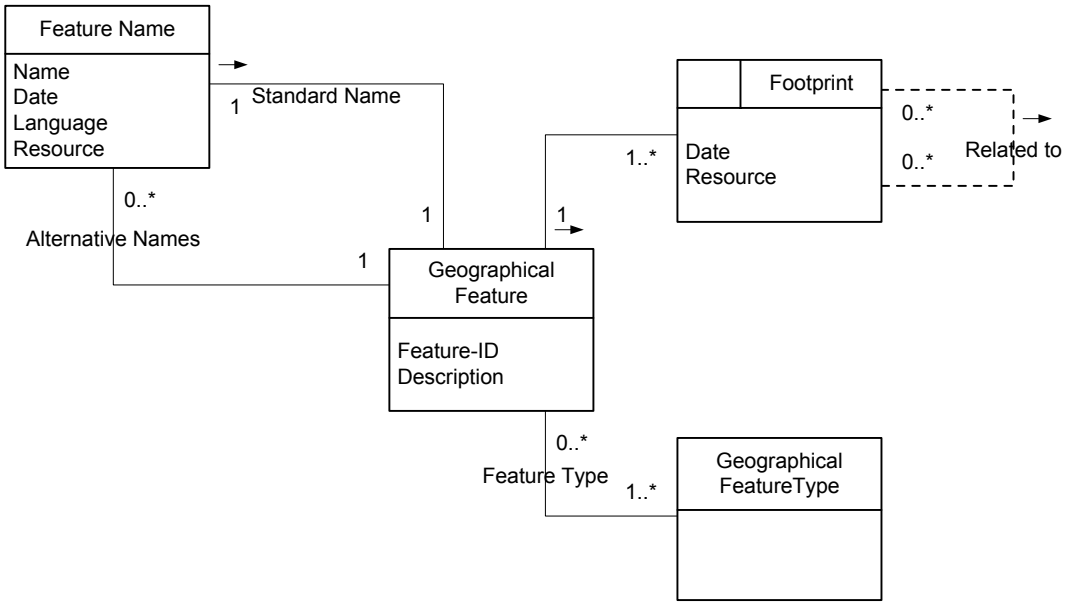


Figura 2.2 – Esquema OMT-G da ontologia de lugar do SPIRIT. Fonte: [16]

Os pesquisadores que desenvolvem a máquina de busca portuguesa Tumba (http://www.tumba.pt), cujo nome é acrônimo de "Temos Um Motor de Busca Alternativo", também têm investigado maneiras de melhorar a qualidade das respostas às consultas geográficas [47]. O processo utilizado é apresentado na Figura 2.3. O PWD (*Purified Web Documents*) é um repositório de páginas que foram coletadas e transformadas em documentos XML bem-formados e que obedecem a um esquema

comum. O texto das páginas é dividido em unidades léxicas, que são comparadas com uma base de conhecimento geográfico para definir o escopo geográfico das páginas. Essa base de conhecimento contém entidades geográficas diversas, como nomes de pontos de referência, endereços, códigos postais e telefônicos. Heurísticas para resolver ambigüidades semânticas dos termos geográficos estão sendo estudadas para tornar essa fase mais eficaz. A etapa seguinte do processo compreende uma análise de *links* entre as páginas, considerando o escopo geográfico (se há uma conexão entre as páginas A e B, e A possui um escopo geográfico definido, então B possui o mesmo escopo). Por fim, uma fase de inferência semântica considera a relação que pode haver entre os diferentes escopos encontrados. O resultado desse processo é chamado de SAWD (*Scope Augmented Web Documents*). Posteriormente, técnicas de indexação agrupam os documentos segundo a similaridade de seus escopos geográficos. Dois documentos são considerados similares se estão relacionados ao mesmo escopo geográfico ou a escopos geograficamente próximos.

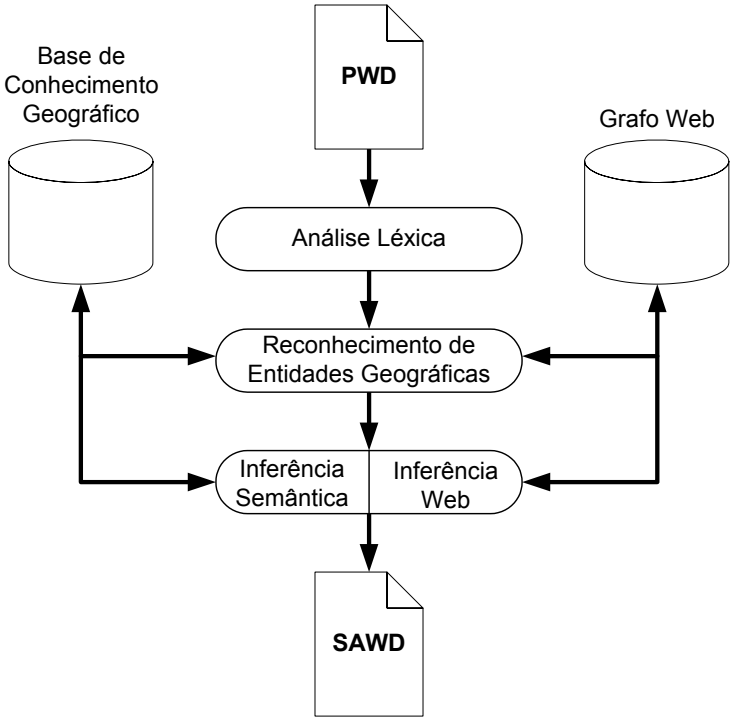


Figura 2.3 – Identificação do contexto geográfico pela Tumba. Fonte: [47]

O Google, máquina de busca de maior utilização no momento, lançou recentemente o Google Local (http://local.google.com). Seu propósito é o de fazer busca em documentos que se referem especificamente a serviços, como, por exemplo, lojas, restaurantes e hotéis. Um catálogo contendo o nome e endereço dos serviços,

semelhante às páginas amarelas, é utilizado na indexação dos documentos. Na interface de consulta, o usuário entra com o serviço e o lugar desejados. A tela de resposta apresenta todos os serviços que atendem à consulta do usuário e um mapa da área em questão. Selecionado um serviço, uma lista dos documentos associados é exibida. Apesar de muito interessante e útil, o Google Local está longe de representar uma solução geral para o problema de recuperação de informação geográfica. A busca é limitada aos serviços contidos nas páginas amarelas e é restrita ao território norte-americano, que contém bancos de dados geográficas confiáveis, atualizadas e de domínio público.

A nova máquina de busca da Microsoft (http://search.msn.com) possuía a opção de busca "*near me*", que apresentava resultados muito semelhantes aos do Google Local. Pouco foi divulgado, porém, sobre o funcionamento da ferramenta, que em 22/04/2005 já se encontrava desativada.

Outras máquinas de busca possuem opções para localização de serviços, assim como o Google Local, porém ainda não fazem busca nos documentos com base na posição geográfica dos serviços localizados. Exemplos são o Yahoo Local (http://local.yahoo.com) e o AOL Local Search (http://localsearch.aol.com).

Zong et al. [54] descrevem um método para atribuição de semântica espacial a páginas Web. A abordagem adotada divide o problema em três etapas: a extração dos nomes de lugares dos documentos, a resolução de ambigüidades e a atribuição dos nomes de lugares às paginas. A extração de nomes de lugares utiliza a ferramenta GATE (http://www.gate.ac.uk) cujo processamento é executado com base em um *gazetteer*. As ambigüidades tratadas são de dois tipos: geo/geo (ambigüidade entre dois nomes de lugares, como município de São Paulo e estado de São Paulo) e geo/não-geo (ambigüidade entre um nome de lugar e um nome que não é um lugar). As ambigüidades são tratadas com um conjunto de regras que consideram a informação de contexto dos documentos e as distâncias euclidianas entre lugares cujos nomes estão sendo analisados. A etapa final consiste na efetiva atribuição de uma localização às páginas. Experimentos utilizando um gazetteer com os dados do censo norte-americano são descritos e indicam resultados eficientes para nomes de lugares com ambigüidade do tipo geo/geo.

O desenvolvimento de métodos que incorporem semântica no processo de busca é um passo importante no sentido de transformar a Web de um repositório de dados em uma grande fonte fornecedora de informações e serviços. Essa é a promessa da Web

Semântica, um ambiente onde os dados seriam mais facilmente explorados, compartilhados e reutilizados pelas aplicações [4]. Em sua vertente geográfica, a Web Semântica Geoespacial [11], a informação geográfica seria capturada, analisada e formatada muito além do nível puramente léxico ou sintático. Os primeiros passos nesse caminho já foram dados, com o surgimento de alguns padrões para a formalização do raciocínio humano, como as linguagens para definição de ontologias (RDF, DAML+OIL e outras), e também o esforço em melhorar a interoperabilidade entre os sistemas de informação geográficos, como a linguagem GML – Geography Markup Language.

Dois requisitos são considerados essenciais para o sucesso da Web Semântica Geoespacial [11]. O primeiro é uma formalização sintática para a elaboração de consultas espaciais. O segundo, mais complexo, é o desenvolvimento de métodos para avaliar a semântica dos recursos disponíveis e julgar se esses recursos podem ser explorados por consultas espaciais. Como proposta para solucionar o primeiro requisito, uma gramática simples, mas que permite a construção de consultas bastante elaboradas, foi introduzida em [11] (Figura 2.4). O símbolo inicial, requisição geoespacial, permite que múltiplas restrições geoespaciais possam ser definidas e combinadas usando operadores lógicos. Uma restrição geoespacial representa uma unidade básica de consulta e é composta por dois termos geoespaciais relacionados por alguma relação espacial (comparador geoespacial). As relações possíveis dependem da ontologia considerada. Um termo geoespacial pode ser uma classe de lugar, também definido em uma ontologia, ou um rótulo (um nome). O uso de um *gazetteer* é sugerido para a definição dos possíveis rótulos.

```
<requisição geoespacial> := <restrição geoespacial>
                            [<operador lógico><requisição geoespacial>]
<restrição geoespacial> := <termo geoespacial>
                            <comparador geoespacial>
                            <termo geoespacial>
<comparador geoespacial> := "baseado em uma ontologia de relações espaciais"
<termo geoespacial> := <classe geoespacial>|<rótulo geoespacial>
<classe geoespacial> := "baseado em uma ontologia de objetos geográficos"
<rótulo geoespacial> := "baseado em um gazetteer geoespacial"
```

Figura 2.4 – Gramática para consultas espaciais.

Em essência, a recuperação de informação geográfica trata da localização de termos de conteúdo geográfico inseridos em texto e da atribuição de uma semântica espacial a

esses fragmentos. Em muitas situações, apenas essa associação não é suficiente. Há circunstâncias em que os fragmentos de texto identificados precisam ser espacialmente localizados, por meio de coordenadas geográficas. A essa atribuição de coordenadas dá-se o nome de *georreferenciamento*. É o georreferenciamento, abordado a seguir, que torna possível, por exemplo, a indexação espacial dos documentos.

2.2 Georreferenciamento

Em SIG, o termo *georreferenciar* diz respeito à tarefa de atribuir coordenadas a uma entidade ou evento espacial, posicionando-a sobre a superfície terrestre com base em um sistema de coordenadas geográficas.

O processo de georreferenciamento pode ocorrer de várias formas. Um receptor GPS pode ser utilizado, por exemplo, para a coleta manual das coordenadas geográficas de um evento. Outra possibilidade é que esse processo ocorra de forma automática a partir de dados alfanuméricos (nomes de lugares, endereços).

O georreferenciamento automático de bancos de dados alfanuméricos é um procedimento útil em muitas situações. Empresas privadas e órgãos públicos possuem cadastros contendo endereços em formato semi-estruturado e que podem ser georreferenciados. O georreferenciamento desses dados pode ajudar enormemente as atividades operacionais e estratégicas dessas instituições. Alguns dos muitos serviços que podem se beneficiar desse processo são o controle da saúde pública [46], o combate à criminalidade urbana [46], o cadastramento escolar [42], a distribuição de mercadorias e a entrega de correspondências.

O georreferenciamento automático de endereços, também chamado de *geocodificação*, é um processo que compreende três etapas [9]: o tratamento do endereço semi-estruturado (*parsing*), o estabelecimento de uma correspondência entre o endereço estruturado e o banco de dados (*matching*) e a atribuição de coordenadas geográficas ao endereço (*locating*). No *parsing*, algoritmos de casamento de padrão podem ser utilizados para gerar uma saída normalizada a partir da entrada semi-estruturada. Esses algoritmos devem conseguir tratar, dentro do possível, inconsistências inevitáveis nessa etapa, já que é muito comum a ocorrência de erros nos endereços semi-estruturados. Para a fase de *matching*, dois tipos de dados são importantes. O principal é a estrutura de endereçamento urbano (endereços individuais, *centerlines* e outras entidades), que consiste em um banco de dados previamente

preparado e que contém o universo dos endereços que poderiam figurar em uma fonte semi-estruturada. Para os casos em que os dados de endereçamento fornecidos não sejam suficientes para a geocodificação e surjam ambigüidades, elementos adicionais podem ser necessários (pontos de referência, nomes e divisas de bairros, códigos de endereçamento postal, e outros), elementos esses também disponíveis em bancos de dados pré-existentes. O resultado final do processo ocorre na fase de *locating*, com a transferência da localização exata ou aproximada de um elemento do banco de dados ao evento que está relacionado ao endereço analisado.

Quando o georreferenciamento automático não tem por objetivo a geocodificação de endereços, o processo se restringe a duas etapas: o reconhecimento do contexto geográfico (*geoparsing*) e atribuição de coordenadas espaciais ao evento (*geocoding*) [33]. Durante o *geoparsing*, palavras-chave do texto que contenham algum significado espacial (nomes de lugares, números de telefone, códigos postais) são identificadas e extraídas. Nessa fase, quase sempre é utilizado um vocabulário controlado de nomes de lugares, geralmente fornecido por um *gazetteer*. No *geocoding*, coordenadas geográficas são atribuídas aos identificadores de lugares reconhecidos na fase anterior.

Em [51], é apresentado o sistema GIPSY, uma ferramenta para o georreferenciamento automático de textos. Inicialmente, a ferramenta executa a análise sintática do texto, identificando termos que casam exata ou aproximadamente com objetos ou atributos de seu tesauro geográfico, que é o nome que usam para seu *gazetteer*. Além disso, construções léxicas (na verdade, relações espaciais) do tipo "adjacente ao lago" e "ao sul do rio" são também identificadas. Na fase seguinte, o sistema seleciona as localizações geográficas que melhor casam com os termos identificados, considerando possíveis modificações devidas às construções léxicas que foram encontradas (por exemplo, "ao sul do rio" provoca um deslocamento da localização geográfica nesse sentido). Uma etapa final tenta inferir qual a melhor localização, analisando a sobreposição das áreas encontradas.

Outros trabalhos existem descrevendo a aplicação de técnicas de aprendizagem de máquina (*machine learning*) para o georreferenciamento de documentos. Em [35], um sistema de localização capaz de gerar mapas com a indicação de lugares citados em notícias de jornal é descrito. A técnica *name entity tagger*, bastante conhecida na área, é utilizada para a identificação dos nomes de lugares, auxiliada por um *gazetteer*. Em [48], o uso de um classificador *Naïve Bayes* para a resolução de ambigüidades que surgem durante o georreferenciamento é descrito.

No entanto, ainda não foi publicada uma solução escalável para o problema de georreferenciamento de texto, assim como ainda não há um padrão para a avaliação da eficiência dos métodos que vem sendo propostos [32].

O consórcio OpenGIS, entidade mais importante no estabelecimento de padrões na área geoespacial, possui duas especificações, em fase de discussão, de serviços acessíveis via rede para georreferenciamento automático. A especificação *Geoparser Service Specification* [39] define um serviço cuja função básica é encontrar elementos que representem lugares de interesse em um texto e retornar a posição da ocorrência desses elementos. A especificação prevê que o usuário seja capaz de indicar quais vocabulários de nomes de lugares deseja utilizar para a execução do serviço. A segunda especificação é a *Geocoder Service Specification* [37], que trata da transformação da descrição textual do lugar (seu nome ou endereço) em uma descrição normalizada, que inclui sua localização geográfica.

O georreferenciamento é um processo sujeito ao surgimento de ambigüidades durante sua execução. Por exemplo, diferentes lugares podem ter um mesmo nome (ex.: "Paris" se refere à capital francesa ou à cidade nos Estados Unidos?). Outro problema é que muitas vezes a entidade alvo do georreferenciamento não aparece indicada pelo seu nome "oficial", mas por um nome conhecido popularmente (ex.: Belo Horizonte, B. H., Belô) ou ainda pelo qual era conhecida em tempos passados. É possível, ainda, inferir o lugar indicado usando diferentes tipos de identificadores, como um CEP ou número de telefone. Percebe-se, portanto, que o georreferenciamento não é uma tarefa trivial, exigindo um tratamento especializado, onde o uso de *gazetteers* digitais pode ajudar enormemente.

2.3 Gazetteers Digitais

Um *gazetteer* é um dicionário geoespacial de nomes geográficos [21]. Trata-se de uma coleção de nomes de lugares acompanhada de sua localização geográfica. Um tipo de *gazetteer* com o qual estamos mais habituados é o índice de um atlas geográfico, onde há uma lista dos nomes dos lugares acompanhado de onde podemos encontrá-los (no caso, a identificação da página do mapa e uma indicação de quadrícula nesse mapa onde o nome poderá ser encontrado).

Mesmo nos países de língua inglesa, onde o conceito é mais difundido e é relativamente comum a publicação de *gazetteers*, não há muito consenso sobre a

utilização do termo [23]. A melhor tradução para o português talvez seja *dicionário toponímico* ou, simplesmente, *dicionário de lugares*.

A função básica de um *gazetteer* é informar a localização de um lugar pelo seu nome. Salvo raras exceções, não utilizamos o formalismo de uma coordenada geográfica quando desejamos nos referir a um lugar. Em nosso discurso falado ou escrito, isto é feito quase sempre utilizando o nome pelo qual o lugar é conhecido, por meio de uma referência espacial indireta [21]. Em várias situações, entretanto, pode ser necessário georreferenciar essa referência espacial indireta, isto é, associar uma ou mais coordenadas geográficas ao nome de lugar. Para essa tarefa, um *gazetteer* é ferramenta fundamental.

Três elementos são considerados essenciais para um *gazetteer* [21]: o nome do lugar, sua localização (*footprint*) e o tipo que identifica a categoria do lugar. O nome pode admitir valores alternativos, como nomes não oficiais, antigos ou em outras línguas. A localização exige ao menos uma coordenada geográfica, mas pode ser também um polígono ou um retângulo envolvente mínimo. O tipo é um atributo fundamental na caracterização do lugar e para solucionar ambigüidades, pois vários lugares podem possuir um mesmo nome.

Com esses três atributos básicos, um *gazetteer* digital é capaz de processar ao menos dois tipos de consultas fundamentais: “onde fica esse lugar?” (ex.: onde fica Belo Horizonte?) e “o que há nesse lugar?” (ex.: que hotéis existem em Belo Horizonte?).

O desenvolvimento de *gazetteers* digitais, em conjunto com o estabelecimento de padrões para o acesso, consulta e compartilhamento de seus dados, pode trazer novas possibilidades para a descoberta e utilização do contexto geográfico em diversas aplicações. O principal papel de um *gazetteer* digital é o de componente na arquitetura de sistemas em que nomes de lugares podem ser empregados para prover algum tipo de informação georreferenciada, como nos sistemas de recuperação de informação geográfica.

Alguns dos *gazetteers* digitais disponíveis na Web estão listados na Tabela 2.1. Pode-se dizer que há muitos outros serviços que não se autodenominam *gazetteers*, mas que possuem um propósito parecido, como páginas amarelas on-line, serviços de localização por mapa e atlas digitais.

Infelizmente, ainda não há padrões estabelecidos para a implementação de *gazetteers*. Cada *gazetteer* listado na Tabela 2.1 possui um esquema próprio de metadados. Além disso, poucos possuem um protocolo de acesso via serviço Web e

aqueles que possuem o implementam cada qual ao seu modo. A falta de padronização dificulta a criação e o uso desses sistemas e cria barreiras para a sua interoperabilidade. Percebe-se, entretanto, o surgimento de algumas iniciativas no sentido de solucionar esse problema.

Nome	Endereço eletrônico
ADL – Alexandria Digital Library Gazetteer	http://middleware.alexandria.ucsb.edu/client/gaz/adl/index.jsp
Australian Government Place Name Search	http://www.ga.gov.au/map/names/
Geographic Names Information System	http://geonames.usgs.gov/gnishome.html
Geographical Names of Canada	http://gnss.nrcan.gc.ca/
GEOnet Names Server (GNS)	http://earth-info.nga.mil/gns/html/
Getty Thesaurus of Geographic Names	http://www.getty.edu/research/conducting_research/vocabularies/tgn/
U.S. Census Bureau	http://www.census.gov/cgi-bin/gazetteer/

Tabela 2.1 – Alguns *gazetteers* disponíveis na Web.

A ISO – *International Organization for Standardization* possui um comitê técnico encarregado de propor padrões na área de informação geográfica digital, o ISO/TC211 (http://www.isotc211.org). O comitê publicou a norma ISO19112 – *Spatial Referencing by Geographic Identifiers*, que define um esquema conceitual para o referenciamento espacial baseado em identificadores geográficos, isto é, sem a utilização de coordenadas espaciais. A especificação, porém, é muito recente (2003) e nenhum dos *gazetteers* pesquisados a implementa. Além disso, trata-se de uma norma comercializada pela ISO, o que impossibilitou-nos o acesso à mesma.

O consórcio OpenGIS possui uma especificação, ainda em fase de discussão, com o título *Gazetteer Service Profile of the Web Feature Service Implementation Specification* [36]. Trata-se, na verdade, de uma extensão da especificação *Web Feature Service* (WFS) [41], um protocolo baseado em HTTP, com requisições e respostas em formato XML, cujo objetivo é permitir que clientes consultem e manipulem objetos geográficos armazenados em um servidor, por meio da linguagem GML [38]. A especificação *Gazetteer Service Profile* adiciona ao WFS alguns recursos específicos para a consulta, inserção e atualização de instâncias armazenadas em *gazetteers* digitais. Estes recursos incluem:

- o acesso aos relacionamentos hierárquicos entre os termos do *gazetteer*, baseado nos conceitos de termo mais geral (*BT – broader term*), termo mais específico (*NT – narrower term*) e termo relacionado (*RT – related term*).

- a recuperação de propriedades específicas de *gazetteers*, tais como o tipo dos lugares.

A versão mais atualizada da proposta data de setembro de 2002. A descontinuidade da discussão talvez se deva ao poder do padrão WFS, que por si só já é suficiente para prover adequadamente a interface básica para um serviço de *gazetteer*.

O *gazetteer* da Alexandria Digital Library (ADL) [1], que pode ser considerado o mais importante disponível no momento, possui e incentiva o uso de seu próprio padrão de metadados para *gazetteers*, o *Gazetteer Content Standard*. Os dados principais que o *Gazetteer Content Standard* admite para o registro de um lugar são:

1. nome, permitindo a inclusão de nomes alternativos;
2. tipo, baseado em uma taxonomia de classificação de lugares bastante abrangente, chamada *Feature Type Thesaurus*;
3. localização pelo retângulo envolvente mínimo ou, quando disponível, pela geometria detalhada do lugar;
4. endereço postal, quando disponível;
5. relacionamento com outros lugares no *gazetteer*, admitindo somente relações topológicas de continência ("parte de") e algumas relações político-administrativas, como "no país".

Além disso, a ADL também desenvolveu um serviço próprio baseado em XML e cujas respostas são registros no formato *Gazetteer Content Standard*.

Segundo [16], os *gazetteers* atuais possuem limitações que impedem sua utilização completa como ferramenta para recuperação de informação geográfica. Em primeiro lugar, os *gazetteers* não tratam de relacionamentos espaciais, a menos de simples relacionamentos hierárquicos. Relacionamentos genéricos entre os objetos também não são implementados, limitando o uso potencial dos *gazetteers* como ontologias geográficas. Além disso, as propriedades dos objetos geográficos não são definidas apenas genericamente e quase sempre há uma perda de características significativos dos objetos. Por fim, os *gazetteers* geralmente contêm nomes associados a objetos bem definidos, e são incapazes de tratar localizações *fuzzy* ou imprecisas, como "sul da Califórnia". A essas observações apontadas por [16], incluímos ainda a falta de nomes intra-urbanos nos *gazetteers*, como pontos de referências, monumentos e outros lugares conhecidos pela população, e também a ausência de alguns tipos de referência importantes, como número de telefone e códigos postais. Nosso objetivo com a

construção do localizador Locus é tentar tratar adequadamente algumas dessas limitações.

3 Projeto do Locus

3.1 Ontologia de Lugar

Segundo a abordagem tradicional de desenvolvimento de sistemas, a primeira etapa de especificação de uma aplicação compreende sua modelagem conceitual. Em contraste com essa abordagem, o desenvolvimento do Locus teve início a partir de uma ontologia que já se encontrava previamente especificada no início dos trabalhos.

O termo ontologia nos foi legado pelos antigos filósofos gregos. Para Aristóteles, uma ontologia era um sistema particular de categorização de certa visão do mundo. Para Gruber, ontologia é "uma especificação explícita de uma conceitualização" [18]. Nesse sentido, uma ontologia é uma descrição dos conceitos e relacionamentos de uma realidade que são relevantes para um agente ou um conjunto de agentes.

O uso de ontologias no desenvolvimento de sistemas de informação tem sido proposto como solução para alguns problemas de inconsistência que ocorrem quando abordagens tradicionais de modelagem são utilizadas [15]. Abordagens tradicionais exigem que o projetista inicie o trabalho de modelagem com a captura da visão de mundo do usuário por meio de um modelo conceitual formal. Essa metodologia obriga o projetista a seguir um paradigma de modelagem, como orientação por objetos ou entidade-relacionamento, o que limita a definição dos conceitos e relacionamentos da realidade aos construtores suportados por esses paradigmas. Esse processo de modelagem introduz inconsistências e incertezas que conduzirão a conflitos inevitáveis entre os conceitos dos usuários e a abstração que foi capturada no esquema conceitual.

Uma diferença fundamental entre ontologias e esquemas conceituais encontra-se nos propósitos aos quais as duas abordagens servem. Enquanto ontologias descrevem domínios específicos do conhecimento, esquemas conceituais são utilizados para descrever a estrutura de bancos de dados [15]. Ontologias são semanticamente mais ricas que esquemas conceituais e, por isso, mais próximas do modelo cognitivo dos usuários. Guarino cunhou a expressão "sistemas de informação dirigidos por ontologia" (*ontology-driven information systems*) para sistemas que fazem uso de ontologias formalmente definidas [19].

Sistemas de informação geográficos são aplicações que podem evoluir muito com o uso de ontologias. A natureza complexa da informação espacial exige representações capazes de descrever não somente atributos dos conceitos, mas também seus componentes geométricos, além de relacionamentos espaciais e temporais. Um modelo conceitual não é capaz de tratar toda essa semântica adicional, característica das aplicações geográficas.

A abordagem tradicional para modelagem de Sistemas de Informação Geográficos é dividida em três níveis [6]. No nível de *representação conceitual*, o desenvolvedor define quais conceitos do mundo real serão representados e também o tipo de representação, entre objetos (entidades individuais e bem definidas) e campos (fenômenos com variação contínua sobre uma superfície). Decidem-se também quais representações geométricas serão utilizadas para capturar cada conceito. Muitas vezes, um conceito pode ter mais de uma representação. Um rio, por exemplo, pode ser representado somente por uma linha ou por um polígono que cobre o espaço entre suas margens. O nível seguinte é o *nível de apresentação*, onde são especificadas características relacionadas com a visualização dos dados, com a definição de parâmetros gráficos, como simbologia e cores. No *nível de implementação*, são escolhidas as estruturas de dados para o armazenamento dos dados espaciais, utilizando ferramentas e linguagens disponíveis nos SIGs ou em SGBDs com recursos geográficos.

Uma variação nesse processo de modelagem é proposta em [15]. Segundo a abordagem proposta (Figura 3.1), ontologias são, inicialmente, utilizadas como ferramenta de formalização dos conceitos e idéias de especialistas que conhecem profundamente o problema modelado. Os projetistas da equipe de desenvolvimento da aplicação definem um esquema conceitual com base nessa ontologia. Eventualmente, os projetistas podem precisar de informações extras dos especialistas, já que as ontologias geralmente representam uma visão de alto nível do problema.

O *gazetteer* do Locus foi modelado com base em uma ontologia de lugar, descrita na Figura 3.2. No início do projeto, essa ontologia já se encontrava formalmente especificada [7] por uma especialista em SIGs urbanos.

Apesar da importância que ontologias vêm adquirindo para representar aspectos da informação, há poucos estudos sobre representação prática do conceito de lugar, especialmente com o propósito de recuperação de informação. Nesta direção, existem alguns trabalhos recentes [26, 27].

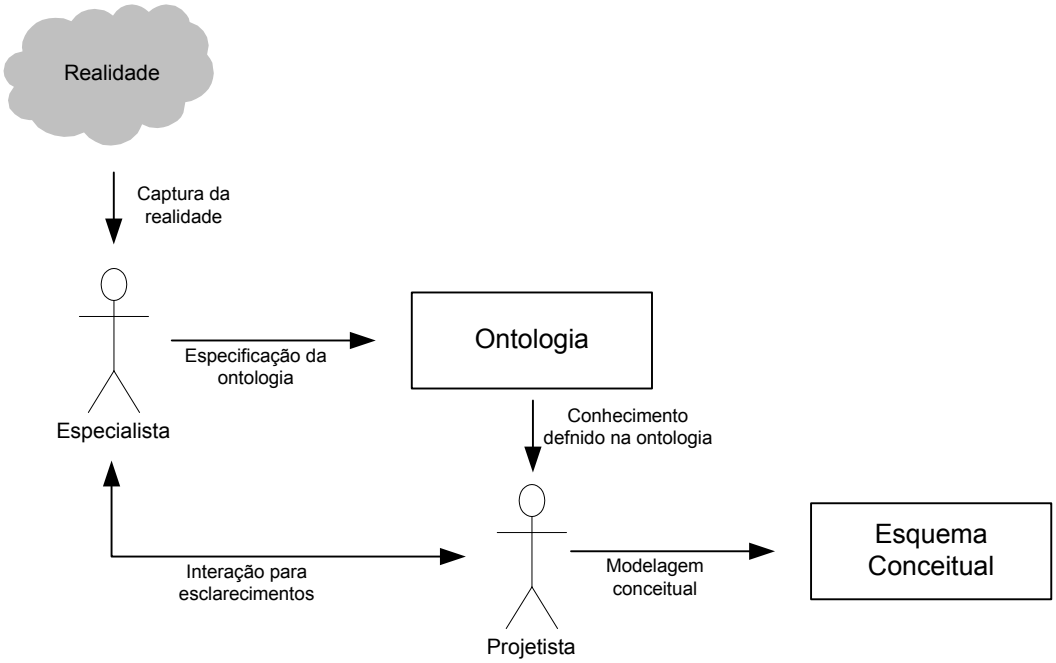


Figura 3.1 – Abordagem dirigida por ontologia para a modelagem de SIGs. Fonte: [15]

A *ontologia de lugar* é uma ontologia particular do espaço geográfico urbano, definida como um conjunto de conceitos dentro desse domínio particular. Essa ontologia descreve feições naturais, objetos ou lugares que possuem significado para uma comunidade urbana, incluindo os relacionamentos entre eles. Um lugar é uma descrição de aspectos do espaço que possui identidade própria. O lugar é mais que uma geometria ou uma topologia, ele inclui um aspecto cognitivo e reflete como as pessoas percebem e usam a informação geográfica.

A ontologia explora a estrutura hierárquica do espaço, onde regiões são subdivididas em outras regiões (subdivisão territorial). Neste domínio, é possível inferir os relacionamentos espaciais, topológicos, de continência e adjacência, utilizando apenas o conhecimento geográfico da estrutura hierárquica. Além da hierarquia de subdivisão territorial, a ontologia ainda inclui conceitos sobre o sistema de endereçamento postal e sobre os locais que constituem pontos de referência para a população.

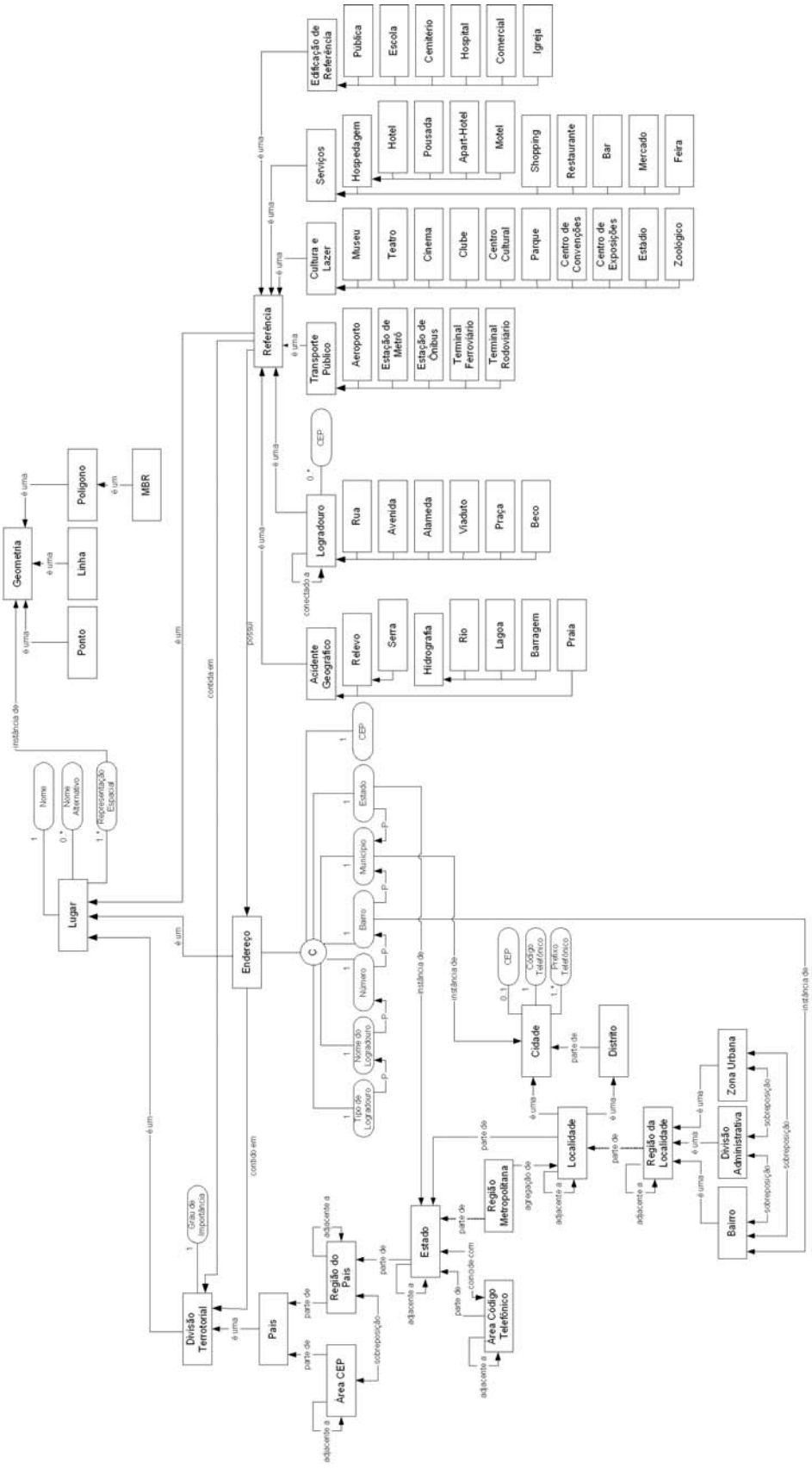


Figura 3.2 – Ontologia de lugar. Fonte: [7]

3.2 Esquema Conceitual OMT-G do Gazetteer

O esquema conceitual do gazetteer do Locus foi especificado usando o modelo OMT-G para aplicações geográficas [6]. Orientado por objetos, o modelo utiliza construtores da linguagem UML e inclui primitivas específicas para representar a geometria e a topologia de dados geográficos. Esse esquema conceitual é mostrado na Figura 3.3. A parte superior do esquema traz a classe **Lugar**, que é a superclasse de todas as entidades espaciais do *gazetteer*. Toda instância dessa classe recebe um identificador único e um nome (com exceção apenas das instâncias da classe **Endereço**). Nomes alternativos ou anteriores podem ser utilizados para identificar um local.

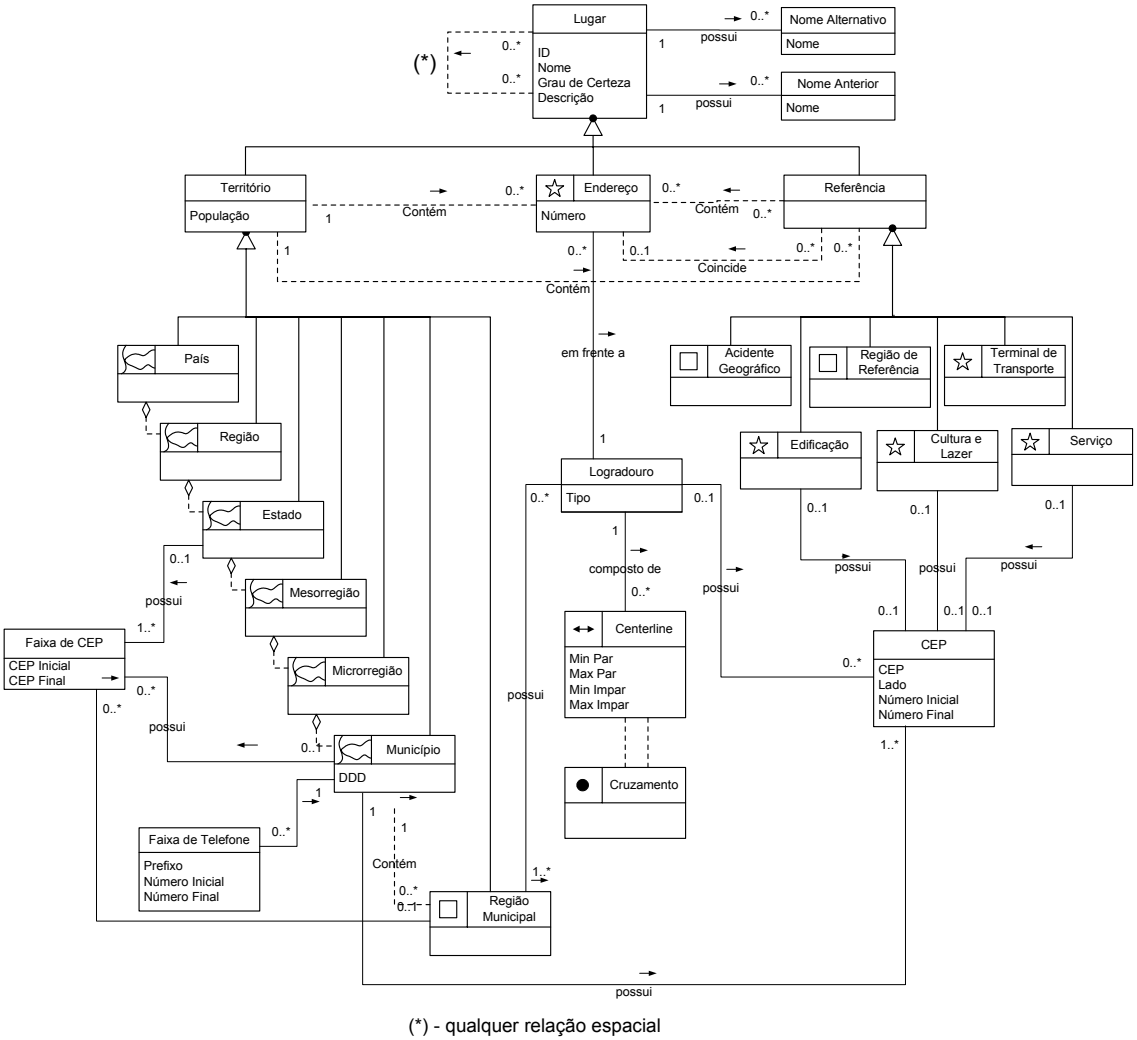


Figura 3.3 – Esquema conceitual OMT-G do *gazetteer* do Locus.

O esquema prevê que várias relações espaciais podem existir entre dois lugares quaisquer. Estas relações podem ser de quatro tipos distintos: topológicas (dentro de,

adjacente a), métricas (a *x* metros de), de ordem (em frente de, ao lado de) ou *fuzzy* (perto de) [12]. Esta forma genérica de representação das relações espaciais tem por objetivo deixar clara a possibilidade de execução de diferentes tipos de consulta pela aplicação, quando essa estiver sendo executada.

A classe **Lugar** se especializa em três subclasses: **Território**, **Endereço** e **Referência**. Os territórios representam as divisões político-administrativas brasileiras. A estrutura hierárquica das subdivisões dos territórios aparece na forma de agregações espaciais. A exceção fica por conta da classe **Região Municipal**, representada por polígonos convencionais, que admitem diferentes tipos de divisões municipais, como bairros, distritos, divisões administrativas e regionais. A adjacência que existe entre as instâncias de **Território** é lembrada com um auto-relacionamento da classe.

Os Correios atribuem faixas de CEP a alguns territórios. Os CEPs do estado de Minas Gerais, por exemplo, estão na faixa entre 30000-000 e 39999-999, e os do município de Belo Horizonte entre 30000-000 e 31999-999. Essa é uma informação importante para associar um número de CEP ao lugar ao qual ele se refere. O esquema admite essa representação na classe **Faixa de CEP**. A classe **CEP** trata daqueles lugares para os quais os Correios atribuem um código de endereçamento único. Isto ocorre para três tipos de lugar: municípios, logradouros de grandes municípios e alguns pontos de referência (grandes usuários dos Correios). No caso dos logradouros, há aqueles para os quais os Correios atribuem mais de um CEP, seccionando o logradouro por faixas de numeração ou pelo lado (lado par e lado ímpar).

Fato semelhante ocorre com o sistema telefônico, no qual são atribuídas faixas de códigos DDD e de números aos municípios e a algumas regiões municipais, especialmente distritos. Por exemplo, uma das faixas de telefone que pertencem ao município de Ouro Preto, MG, possui DDD 31, prefixo 3551 e números entre 0000 e 6199. Esse conceito está representado no esquema pela classe **Faixa de Telefone**.

Os objetos da classe **Endereço** correspondem aos endereços urbanos com localização conhecida. Estes objetos são armazenados individualmente (um ponto para cada endereço). A opção pelo registro individual simplifica a representação e consegue tratar adequadamente problemas que são comuns em cidades brasileiras, como logradouros que possuem numeração irregular. O esquema prevê a possibilidade do armazenamento da estrutura viária básica do município, que inclui as *centerlines* (eixos de vias) e os cruzamentos. Opcionalmente, caso não exista informação de

endereçamento individual georreferenciado, faixas de numeração podem ser associadas aos *centerlines*.

Lugares conhecidos da população e utilizados como pontos de referência são representados na classe **Referência**, que se especializa em classes mais representativas do tipo do ponto de referência, como **Acidente Geográfico**, **Terminal de Transporte**, e **Cultura e Lazer**.

Terminada a modelagem conceitual do *gazetteer*, foi criado um esquema lógico, para implementação em um SGBD relacional (Figura 3.4). Uma ferramenta CASE capaz de gerar o DDL SQL do esquema de forma automática foi empregada nessa etapa.

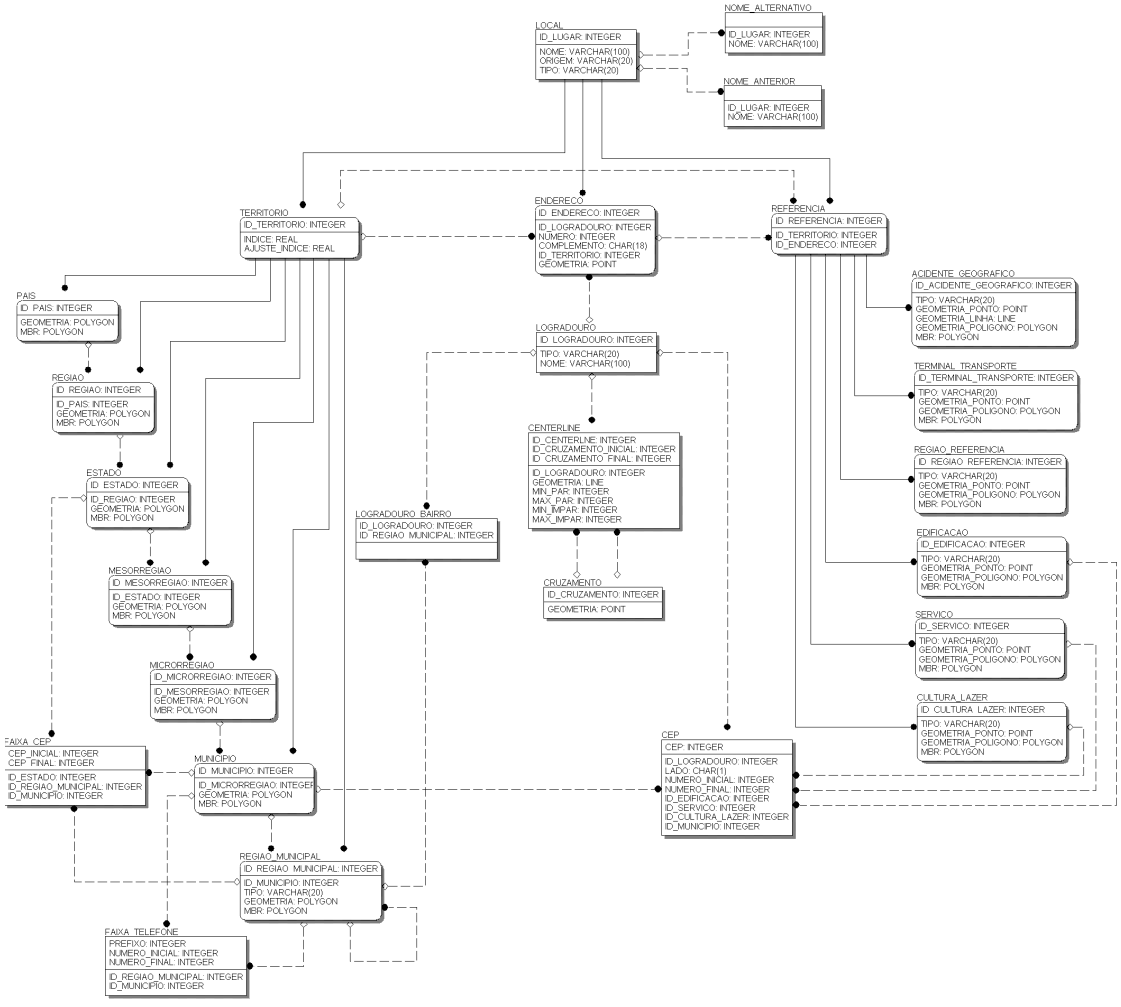


Figura 3.4 – Esquema lógico do *gazetteer*.

3.3 Modelagem do Locus

Em paralelo ao trabalho de modelagem do *gazetteer*, ocorreu o trabalho de especificação funcional e implementação do sistema. É importante mencionar que o Locus não é um *gazetteer*, mas sim um sistema que se baseia em um g*azetteer* para prover as funcionalidades que são o seu objetivo. Logo, foi necessário implementar o sistema que, em conjunto com o *gazetteer*, compõe a solução.

Alguns requisitos básicos foram definidos ainda no início do projeto [50]:

- acesso via Web;
- consultas capazes de fazer busca aproximada, admitindo erros na entrada do usuário;
- minimização do tempo de resposta às consultas;
- navegação baseada na ontologia, isto é, a capacidade de adaptar a interface de resposta da consulta ao tipo do lugar localizado, incluindo opções para a expansão da consulta, como, por exemplo, encontrar lugares próximos ou o território onde o lugar está contido.

A linguagem de modelagem UML [5] foi utilizada para a especificação do sistema. O diagrama de casos de uso aparece na Figura 3.5.

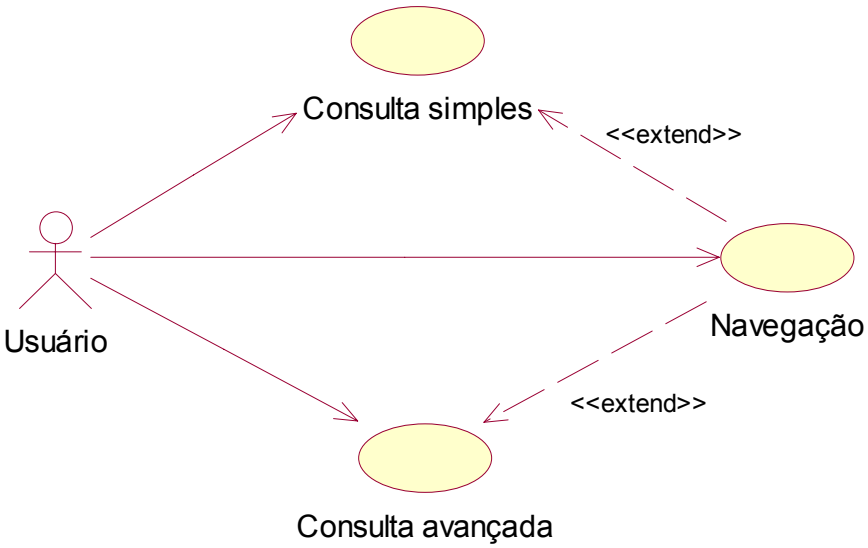


Figura 3.5 – Diagrama de casos de uso.

Dois tipos de consulta são possíveis. No caso de uso *Consulta simples*, o usuário entra com o texto para busca e, em seguida, o sistema localiza todos os registros cujos nomes casam exata ou aproximadamente com a entrada do usuário (procedimento

descrito na seção 4.3) . Para o caso de mais de um registro ter sido localizado, esses são ordenados segundo um *ranking* de relevância de lugares, de forma que os mais importantes apareçam nas primeiras posições (procedimento descrito na seção 4.4). Na seqüência, o usuário escolhe o registro que lhe interessa e o sistema mostra todas as informações disponíveis no *gazetteer* sobre o lugar, incluindo um mapa de sua localização. O diagrama de seqüência do caso de uso *Consulta simples* é exibido na Figura 3.6.

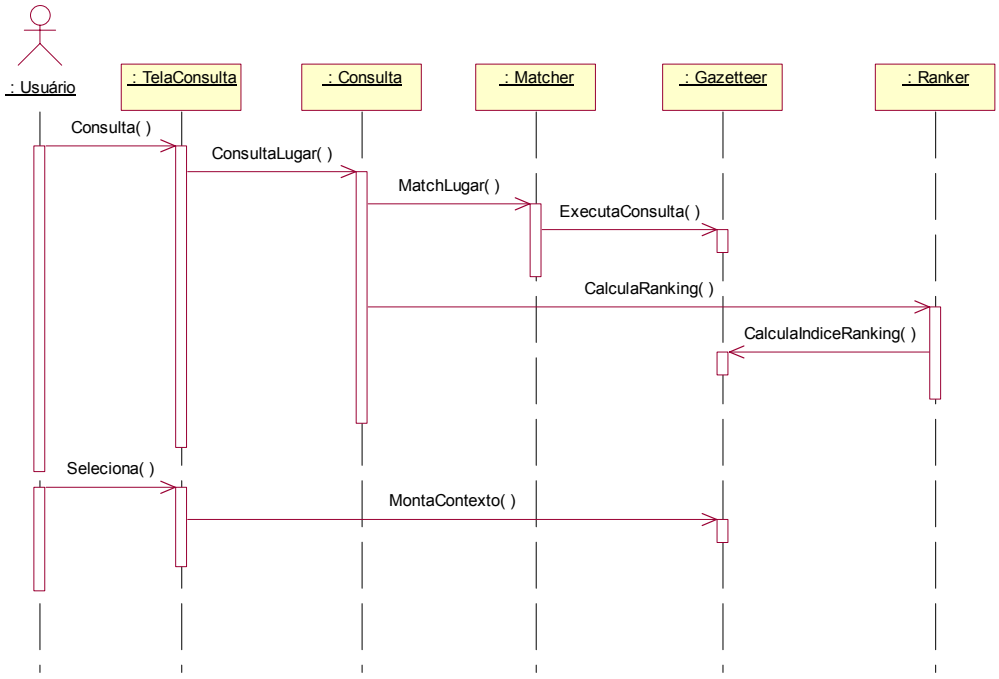


Figura 3.6 – Diagrama de seqüência do caso de uso Consulta simples.

Já no caso de uso *Consulta avançada*, é possível utilizar relações espaciais para filtrar o resultado da pesquisa. O usuário entra com uma trinca de valores, no formato definido pela gramática para consultas descrita na Figura 2.4:

<termo geoespacial> <comparador geoespacial> <termo geoespacial>

Os termos geoespaciais são tipos ou nomes de lugares, enquanto o comparador geoespacial representa uma relação espacial. Ao primeiro termo geoespacial da consulta deu-se o nome de "lugar de interesse" e ao segundo "lugar de referência". Essa estrutura permite a construção de buscas elaboradas, como, por exemplo, "restaurante" "perto da" "Praça da Liberdade" e "hotel" "a 1.000 metros do" "Mineirão". O diagrama de seqüência desse caso de uso é exibido na Figura 3.7. Primeiramente, é feita uma busca para localizar o lugar de referência no *gazetteer*. Esse procedimento é idêntico à execução de uma consulta simples. Em seguida, os lugares de interesse são localizados

e a consulta espacial é elaborada e executada, filtrando os lugares de interesse que a ela atendem.

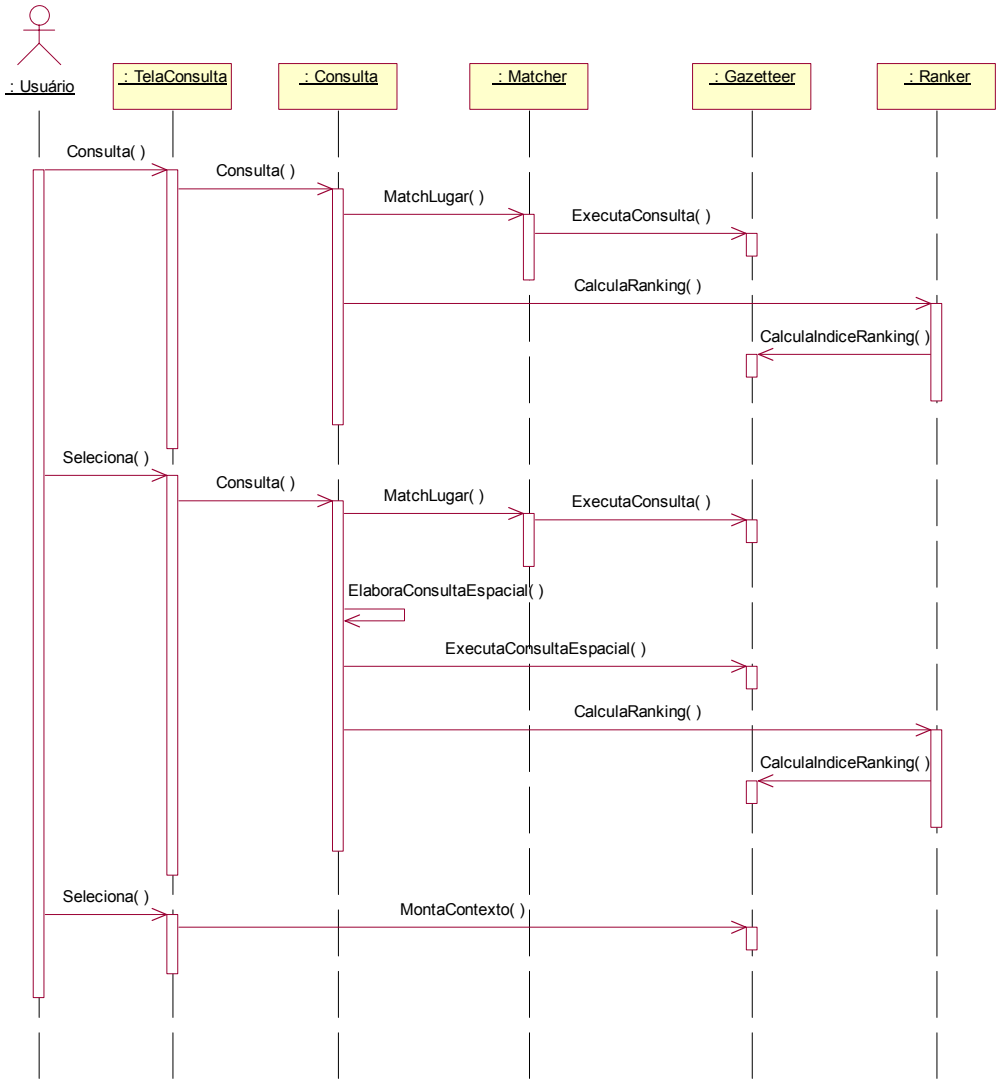


Figura 3.7 – Diagrama de seqüência do caso de uso Consulta avançada.

O caso de uso *Navegação* estende a funcionalidade dos casos de uso *Consulta simples* e *Consulta avançada*, prevendo a situação em que, terminada a busca, o usuário navega pela ontologia do *gazetteer*. Isso é possível com o a exibição, na tela de resposta, de um conjunto de opções para a pesquisa de lugares relacionados geograficamente ao lugar encontrado.

A arquitetura do sistema pode ser observada na Figura 3.8. O componente *Consulta* possui um papel central como controlador da lógica da aplicação, coordenando a operação dos demais componentes. O *Matcher* encapsula os algoritmos de busca exata e aproximada e o *Ranker*, a lógica de elaboração do *ranking* de relevância de lugares. O

Servidor de Mapas é encarregado da montagem e exibição dos mapas. Um *Servidor WFS* também está incorporado ao sistema, permitindo consultas por meio desse protocolo.

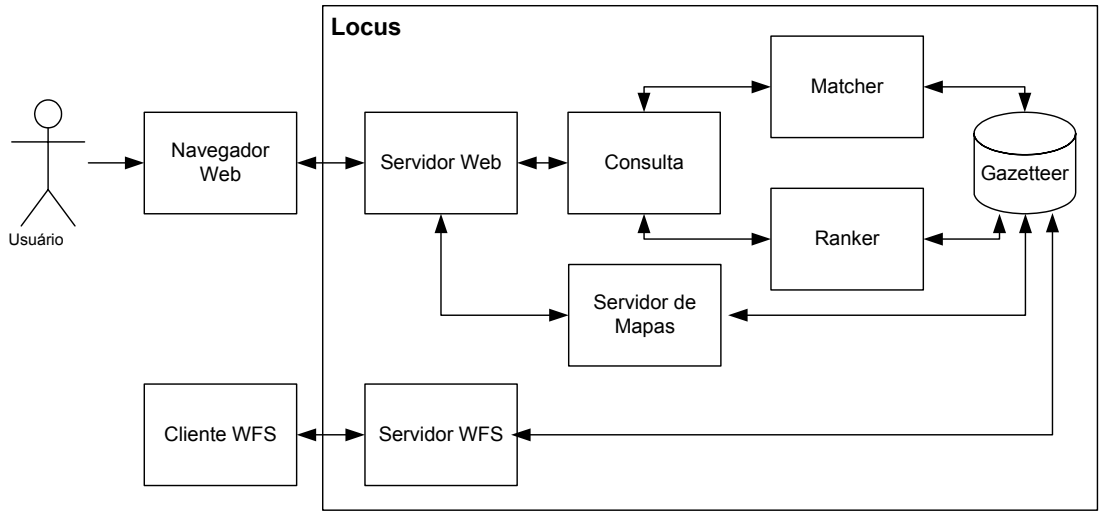


Figura 3.8 – Arquitetura do sistema.

4 Implementação e Avaliação do Locus

4.1 Opções de implementação

O Locus foi implementado na linguagem de programação JAVA (http://java.sun.com). Essa escolha ocorreu principalmente porque os recursos disponíveis atualmente implementados na linguagem atendem de maneira suficiente aos requisitos do sistema, especialmente no que diz respeito ao desenvolvimento de aplicações Web. Atualmente, muitas ferramentas e bibliotecas de software baseadas em JAVA e no *framework* de desenvolvimento J2EE encontram-se disponíveis sob a licença de software livre. O uso desses recursos traz grandes vantagens, como a melhoria na qualidade dos programas criados e a diminuição no tempo de desenvolvimento de novas aplicações.

Para a implementação das páginas de conteúdo dinâmico da aplicação, foi utilizada a tecnologia de *servlets* e JSP (*Java Server Pages*). O *servlet container* (o programa encarregado de executar os *servlets*) escolhido foi o Apache Tomcat.

Outra decisão importante foi quanto à escolha de um SGBD com capacidade de armazenamento de dados espaciais. A opção recaiu sobre o PostgreSQL (http://www.postgresql.org), um SGBD objeto-relacional distribuído sob a licença de software livre e que possui um módulo específico para o tratamento de dados espaciais, o PostGIS (http://postgis.refractions.net). Este módulo implementa o padrão *Simple Feature Specifications* for SQL do consórcio OpenGIS [40], que especifica uma extensão do padrão SQL92 para permitir o armazenamento, recuperação, consulta e modificação de objetos geográficos em SGBDs relacionais. A especificação inclui um conjunto de tipos geométricos válidos, além de funções espaciais para sua manipulação. As funções espaciais utilizadas pelo PostGIS são implementadas pela biblioteca GEOS (http://geos.refractions.net) e os cálculos de projeção cartográfica pela biblioteca Proj4 (http://www.remotesensing.org/proj). O algoritmo de indexação espacial utilizado é o GiST (http://www.sai.msu.su/~megera/postgres/gist), que é uma variação das árvores-R [20]. Todos esses projetos são desenvolvidos pela comunidade de software livre e distribuídos gratuitamente.

O servidor de mapas escolhido foi o MapServer (http://mapserver.gis.umn.edu). Servidores de mapas são aplicações utilizadas para a criação e distribuição de mapas estáticos ou dinâmicos via Web. O MapServer foi escolhido principalmente pela capacidade de conseguir acessar diretamente dados espaciais armazenados no PostGIS.

Uma forma alternativa de acesso ao Locus é através do protocolo Web Feature Sevice (WFS), que fornece uma interface com operações de consulta e manipulação de objetos geográficos sobre HTTP e em formato XML. Para disponibilizar esse serviço no Locus, optou-se por utilizar o servidor WFS GeoServer (http://geoserver.sourceforge.net).

4.2 Carga de Dados

Antes de iniciar a inserção de dados, um sistema de coordenadas geográficas único foi escolhido para o armazenamento de todos os objetos espaciais no *gazetteer* do Locus. O objetivo principal dessa decisão foi o de evitar o custo adicional e desnecessário do processamento de conversões de projeções *on-line*.

Dado o caráter universal que se deseja para um *gazetteer*, um sistema de coordenadas ideal seria um que fosse válido para todo o planeta. Tal sistema, porém, não existe. No caso do Locus, como a cobertura dos dados corresponde à área do território brasileiro, decidiu-se pela utilização do sistema de coordenadas geográficas (latitude e longitude) expresso em graus, com datum SAD69. Essa decisão exigiu que, durante a carga de dados, algumas conversões fossem executadas, geralmente do sistema UTM para o sistema de coordenadas geográficas escolhido. Essas conversões foram executados com a ferramenta Projection Utility do SIG ArcView 3.2.

É importante mencionar que nem toda entrada no *gazetteer* do Locus possui obrigatoriamente uma localização espacial explicitamente registrada. Muitos logradouros, por exemplo, não estão associados a um conjunto de *centerlines* que os localizam, assim como muitas referências não possuem sua localização exata. Para esses casos, a localização mais precisa relacionada ao lugar é sempre retornada nas consultas. Assim, se o *gazetteer* não possui a localização exata do logradouro, mas possui a localização do bairro, esta é retornada como a localização do logradouro. Para todo lugar intra-urbano, é sempre garantido que ao menos o retângulo envolvente mínimo do município onde o lugar se encontra é retornado na resposta das consultas. A qualidade

da resposta do *gazetteer* depende, portanto, da qualidade dos dados armazenados. É possível avaliar quantitativamente essa qualidade pelos critérios definidos em [9].

Os dados inseridos no *gazetteer* na primeira carga executada aparecem listados na Tabela 4.1. A malha municipal brasileira do IBGE, atualizada até ano de 2001, forneceu a divisão territorial brasileira até o nível de município. Constam desta base os nomes e as geometrias de contorno do país, das regiões, dos estados, das mesorregiões, das microrregiões e dos municípios, além de vários atributos relativos ao Censo de 2000.

Tipo	Quantidade	Fontes
Países	1	IBGE
Regiões	5	IBGE
Estados	27	IBGE
Mesorregiões	137	IBGE
Microrregiões	558	IBGE
Municípios	5.560	IBGE
Regiões Municipais	40.835	IBGE, Correios, ANATEL e PRODABEL
Referências	10.400	Correios
Logradouros	596.312	Correios e PRODABEL
Endereços	420.942	PRODABEL

Tabela 4.1 – Carga inicial de dados no *gazetteer* do Locus.

Outra fonte de dados importante foi o catálogo de CEP dos Correios. Os logradouros das 346 maiores cidades brasileiras possuem CEPs individuais. O acesso a esses dados permitiu a inclusão do nome de uma grande quantidade de bairros e de logradouros desses municípios. O catálogo de CEP dos Correios também forneceu o nome de mais de 10 mil referências que representam seus grandes usuários (como edifícios, universidades e hospitais) e que possuem um código de CEP único.

Dos registros dos códigos de área telefônicos da Agência Nacional de Telecomunicações (ANATEL), foram extraídos nomes de distritos dos municípios de todo o país, pois muitos distritos possuem prefixos telefônicos exclusivos.

Para o município de Belo Horizonte, foi possível contar com a malha viária urbana e também com os endereços individuais fornecidos pela Empresa de Informática e Informação do Município de Belo Horizonte (PRODABEL), além de nomes de bairros e alguns pontos de referência.

Os dados dos Correios e da ANATEL também foram úteis, claro, não só para o fornecimento dos nomes de lugares, mas também como fontes dos números de CEP e de códigos de área telefônicos de todo o país. O *gazetteer* foi alimentado com 621.293 números de CEP e 81.880 códigos de área telefônicos (combinações únicas de DDD e prefixos telefônicos).

Uma dificuldade encontrada nesta etapa do trabalho foi a falta de uniformidade no registro dos nomes dos lugares nas principais fontes de dados utilizadas (IBGE, Correios e ANATEL). Esse problema ocorreu principalmente no que se refere aos nomes dos municípios. Muitos municípios cujos nomes constavam na base de um desses órgãos não apareciam nos demais. Esse problema ocorria principalmente devido à ocorrência de emancipações, além de algumas modificações nos nomes dos municípios. Um trabalho de uniformização dos nomes dos lugares foi necessário para integrar essas bases. Nesse procedimento, a base do IBGE foi considerada padrão e os nomes dos municípios nas bases dos Correios e da ANATEL foram modificados para se adequarem aos registros do IBGE. A consulta às páginas na Web dos municípios que apresentaram problemas foi a principal forma de solucionar dúvidas. Por fim, 6 municípios que constavam da base dos Correios ficaram sem correspondente no IBGE e, por isso, não foram inseridos no *gazetteer*. Como se trata de um universo de 5.560 municípios, essas ausências podem ser consideradas pequenas.

A inserção dos dados espaciais no PostGIS foi executada com a ferramenta shp2pgsql, que acompanha o PostGIS e converte arquivos no formato *shape* do ArcView em scripts SQL, que posteriormente são executados no SGBD.

Ao fim dessa primeira fase de alimentação, percebeu-se que o *gazetteer*, apesar de já conter um volume significativo de dados, ainda não possuía uma quantidade representativa de pontos de referência, especialmente lugares intra-urbanos como serviços (hotéis, restaurantes, teatros, museus e outros). Uma nova carga de dados, seguindo uma nova abordagem, foi então executada. Esta nova etapa consistiu na coleta de páginas Web com referências a serviços, seguida da extração dessas referências e de sua inserção no *gazetteer*.

A tarefa de extrair dados de interesse de um sítio Web é executada por programas chamados *wrappers* [29]. Estes programas são responsáveis por coletar as páginas, identificar e extrair os dados desejados das páginas coletadas e armazenar os objetos extraídos em um formato adequado. O desafio maior de um *wrapper* é ser capaz de reconhecer os dados de interesse em meio a outros elementos de texto que não interessam para a extração.

A geração de *wrappers* é um problema que pode ser definido como se segue. Dada uma página Web *S* contendo um conjunto implícito de objetos, determine um mapeamento *W* que alimenta um repositório de dados *R* com os objetos em *S*. O mapeamento *W* precisa também ser capaz de reconhecer e extrair dados de qualquer

página *S'* similar a *S* [29]. De uma maneira geral, o mapeamento *W* é um conjunto de regras ou padrões de texto que reconhece valores de atributos para os objetos extraídos.

Desenvolvido no Laboratório de Bancos de Dados da UFMG, o ambiente WByE (*Wrapping By Example*) [17] é capaz de gerar *wrappers* com base em exemplos fornecidos pelo usuário. Dos exemplos fornecidos, são geradas especificações, que por sua vez são utilizadas para a criação de agentes capazes de coletar e extrair dados dos sítios Web.

O ambiente WByE é composto por duas ferramentas integradas: ASByE (*Agent Specification by Example*) e DEByE (*Data Extraction by Example*). A primeira é utilizada para gerar um plano de coleta de páginas (PFP - *page fetching plan*) que guia o comportamento de um agente responsável pela coleta do conjunto de páginas do sítio Web. O usuário interage com a ASByE por meio de uma interface gráfica (Figura 4.1) que usa uma estrutura de grafo para representar uma porção da Web. Assim, o usuário navega pelos vértices do grafo (as páginas) explorando as várias arestas disponíveis entre eles (os *links* entre as páginas). Essa navegação fornece exemplos para a ASByE de como alcançar as páginas desejadas, como preencher formulários necessários e como navegar por uma coleção de páginas relacionadas.

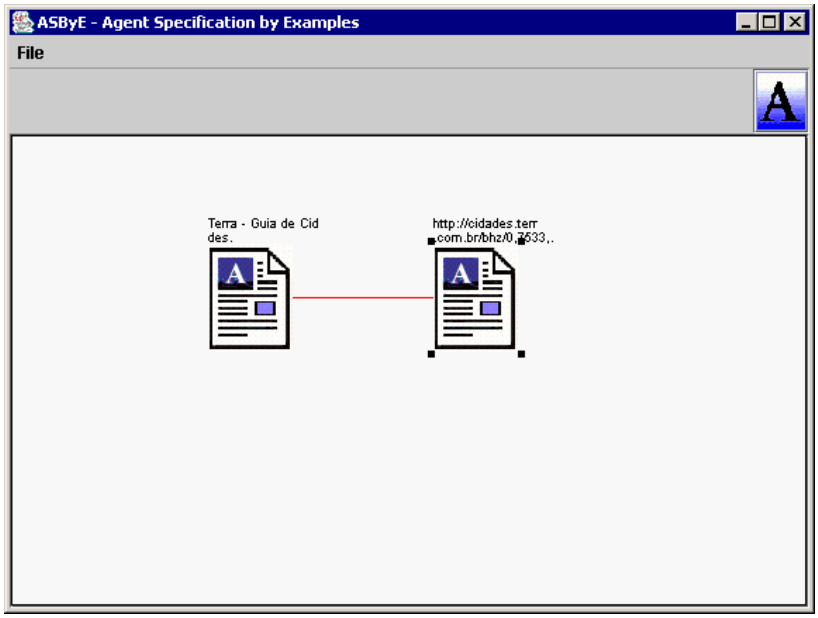


Figura 4.1 – Interface gráfica da ASByE.

A segunda ferramenta, DEByE, é utilizada para especificar como os dados devem ser extraídos das páginas coletadas e como organizá-los logicamente de acordo com a percepção do usuário da estrutura implícita dos dados nas páginas [30]. DEByE captura essa percepção utilizando uma forma estendida do conceito de tabelas aninhadas, onde

uma coluna da tabela pode ter duas ou mais sub-estruturas diferentes. A interface da ferramenta (Figura 4.2) permite a especificação das tabelas aninhadas, cridas pelo usuário em um procedimento de copiar e colar pedaços de dados contidos nas páginas fornecidas como exemplo. Dos exemplos fornecidos, a ferramenta gera um padrão de extração de objetos (OEP - *object extraction pattern*), que descreve a estrutura dos objetos a serem extraídos e o contexto capaz de caracterizar os atributos atômicos dos objetos no texto (como marcações HTML, símbolos e palavras-chave).

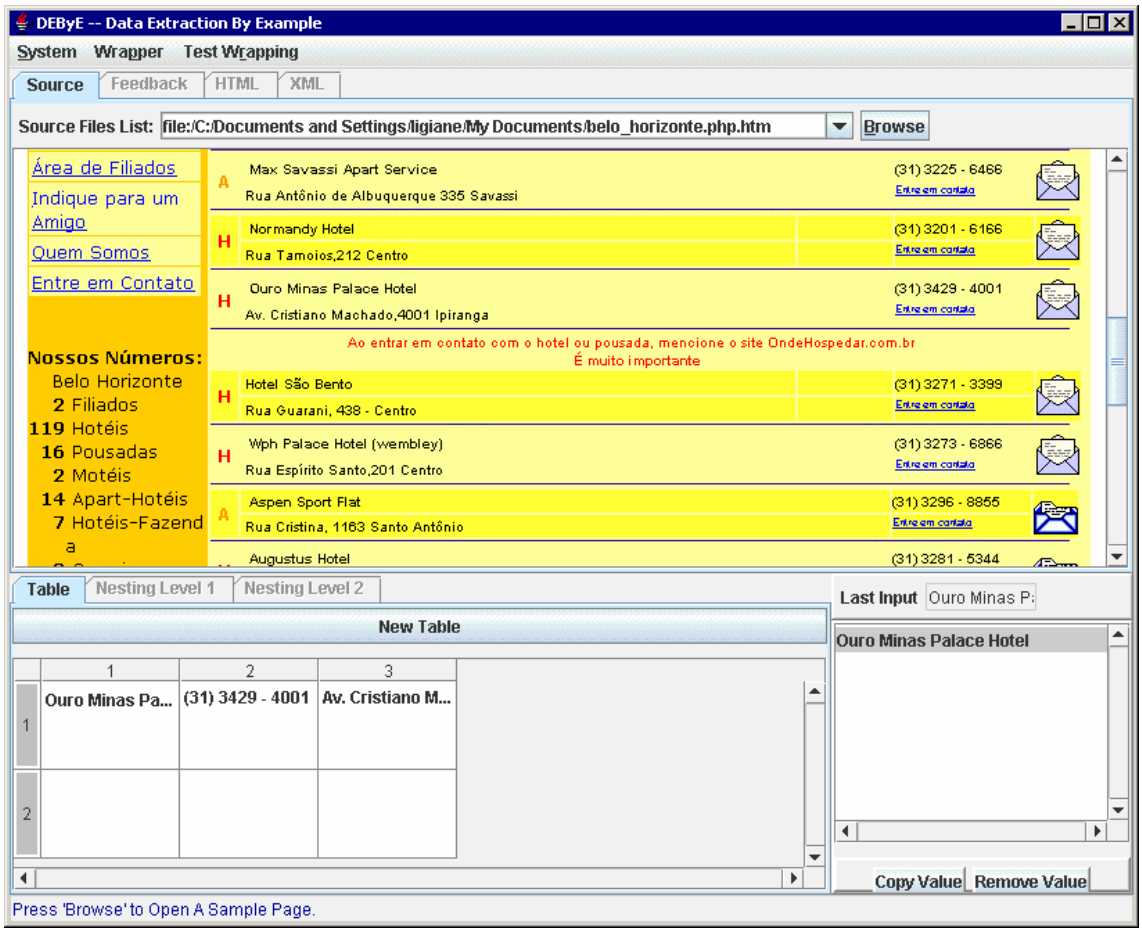


Figura 4.2 – Interface gráfica da DEByE.

A inserção dos registros extraídos no *gazetteer* exigiu ainda a implementação de um módulo de análise de similaridade, para evitar que uma referência já armazenada fosse inserida novamente. A Figura 4.3 traz o código em alto nível do algoritmo para a determinação da similaridade entre duas referências. O atributo telefone foi considerado um caracterizador inequívoco de similaridade positiva. Para os casos em que as referências não possuem telefone, a similaridade é verificada pela combinação do nome e endereço ou do nome e tipo das referências comparadas. A comparação de nomes é feita pelo cálculo da distância de edição de Levenshtein (número mínimo de operações

de inserção, remoção e substituição de caracteres para tornar duas cadeias iguais [34]). A comparação de endereços é baseada na identificação do nome da rua e do número nos dois endereços, sendo que os endereços comparados são considerados iguais se a distância de edição entre os nomes das ruas atender a um limite (que definimos em 20%) e os números forem iguais.

```
Testa Similaridade (ref1, ref2)
begin
   if ref1(cidade, estado) ≠ ref2(cidade, estado)
      return false;
   if ref1(telefone) = ref2(telefone)
     return true;
   if ref1(endereço) = ref2(endereço)
      if DistânciaEdição (ref1(nome), ref2(nome)) < α
        return true;
      if ref1(nome) contido em ref2(nome) ou ref2(nome) contido em
ref1(nome)
        return true;
   if DistânciaEdição (ref1(nome), ref2(nome)) < α
      if ref1(tipo) = ref2(tipo)
        return true;
   return false;
end;
```

Figura 4.3 – Algoritmo para determinação da similaridade entre duas referências.

Um diagrama descrevendo todo o processo de alimentação do *gazetteer* pela Web pode ser observado na Figura 4.4.

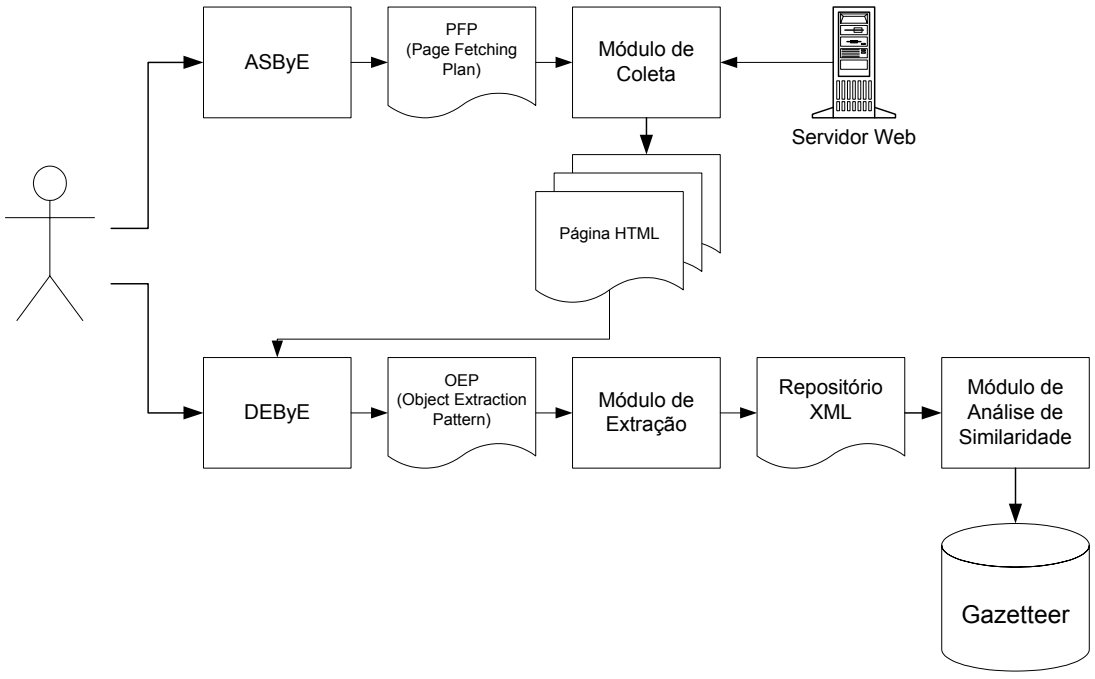


Figura 4.4 – Processo de alimentação do *gazetteer* pela Web.

Para a coleta das páginas, foram selecionados sítios da Web que pudessem fornecer ao menos três atributos básicos para uma caracterização mínima das referências: *nome*, *município* e *estado*. Outros três atributos opcionais também foram extraídos quando estavam disponíveis: *tipo*, *endereço* e *telefone*. Quando não era possível definir o tipo da referência, essa era classificada simplesmente como "ponto turístico".

Seis sítios da Web tiveram suas páginas coletadas, como pode ser observado na Tabela 4.2. Em alguns desses sítios, o agente gerado pela ASByE teve que ser adaptado para coletar páginas cuja URL deveria conter o nome da cidade das referências (por exemplo, http://www.citybrazil.com.br/mg/belohorizonte/turismo.htm). A adaptação consistiu em fazer o agente ler um arquivo contendo a lista dos municípios brasileiros para gerar as possíveis URLs. Com o objetivo de aumentar a taxa de acerto, foram inseridas nesse arquivo algumas variações nos nomes dos municípios que incluíam, por exemplo, nomes sem preposição ou com possíveis abreviaturas. No total, foram coletadas 7.741 páginas e extraídas 33.315 referências.

Endereço	Abrangência	Conteúdo	Páginas coletadas	Referências extraídas
cidades.terra.com.br	9 capitais e outras 13 cidades	restaurantes, bares, casas noturnas, teatros, museus e outros	278	2.443
www.cidades.com.br	todo o país	praias, hotéis e restaurantes	1.416	10.588
www.guiadasemana.com.br	7 capitais	restaurantes, bares, pontos turísticos, casas noturnas e outros	15	2.977
www.citybrazil.com.br	todo o país	pontos turísticos	4.900	4.628
www.ondehospedar.com.br	todo o país	hotéis	1.131	12.286
www.pachecodrogaria.com.br	2 estados (MG e RJ)	hospitais e postos de saúde	1	393

Tabela 4.2 – Resultados da coleta de páginas e da extração de referências.

O procedimento de inserção das referências no *gazetteer* consistiu em verificar, para toda referência candidata a inserção, se havia alguma referência similar no conjunto já inserido no *gazetteer*. Se negativo, a inserção da nova referência era executada. Se positivo, os dados da referência similar já armazenada eram complementados com os dados que a referência candidata a inserção possuía e eram ausentes na referência armazenada. Se, por exemplo, a referência candidata a inserção trazia um número de

telefone que a referência armazenada não possuía, este era copiado para a referência armazenada.

Ao fim do procedimento, 29.139 referências foram inseridas no *gazetteer* e 1.715 referências já armazenadas sofreram um ganho de informação (inserção do tipo, endereço ou telefone da referência).

4.3 Implementação do Módulo de Busca

As consultas ao Locus ficariam muito limitadas se exigissem do usuário a entrada do nome exato dos lugares para a busca. Uma funcionalidade desejada para o sistema desde o início do projeto é justamente a capacidade de executar busca aproximada, admitindo erros.

O problema que se deseja resolver é o seguinte: dadas duas cadeias de caracteres (a digitada pelo usuário e cada um dos nomes armazenados no *gazetteer*), como determinar o quão semelhantes elas são? O método mais utilizado para determinar essa semelhança consiste no cálculo da distância de edição de Levenshtein, que corresponde ao número mínimo de operações de remoção, inserção e substituição necessárias para tornar as duas cadeias idênticas [34]. O algoritmo tradicional para o cálculo da distância de edição de Levenshtein data dos anos 60 e é baseado na técnica de programação dinâmica. A complexidade de tempo do algoritmo é $O(mn)$.

Um problema relacionado pode ser encontrado na literatura com o nome de casamento de padrão (*pattern matching*) [53], e é formalizado como se segue. Dado um texto $T[1..n]$ de tamanho n e um padrão $P[1..m]$ de tamanho $m \leq n$, onde os elementos de P e T são escolhidos de um alfabeto finito Σ de tamanho c, determinar as ocorrências de P em T . O algoritmo Shift-And [52] implementa uma solução elegante e eficiente (complexidade do pior caso $O(kn)$) para o problema de casamento aproximado de padrão, modelando a pesquisa como um autômato finito não-determinístico.

Nas primeiras implementações, o processamento de busca no *gazetteer* consistia em percorrer todos os registros de nomes de lugar no SGBD e calcular a distância de edição de Levenshtein de todos eles em relação ao nome informado pelo usuário. Os registros recuperados eram, então, ordenados segundo a distância de edição e o índice calculado para o lugar no *ranking* de lugares (vide seção 4.4). Esse procedimento demonstrou ser muito ineficiente, chegando a demorar vários minutos para sua conclusão. A lentidão

devia-se principalmente à complexidade quadrática do algoritmo de cálculo da distância de edição utilizado.

A solução encontrada para tornar a busca mais rápida consistiu em dividir o procedimento em duas etapas: na primeira, o algoritmo Shift-And, bem mais eficiente, é utilizado para indicar se a cadeia de caracteres de pesquisa (o nome digitado pelo usuário) ocorre nas cadeias pesquisadas (os nomes armazenados no SGBD) com até *k* erros. Somente quando essa verificação é positiva, a distância de edição entre as cadeias de caracteres é calculada. Um casamento ocorre se a distância de edição calculada for inferior a um limite α previamente estabelecido.

Essa estratégia introduziu um ganho de desempenho considerável. Porém, o tempo necessário para processar a consulta, através da interface JDBC de acesso ao banco de dados, para retornar os quase um milhão de nomes armazenados consumia muito tempo e memória, fazendo até com que a aplicação não conseguisse responder a algumas consultas. Esse problema foi solucionado pela implementação de um programa em linguagem C para executar o algoritmo Shift-And diretamente sobre um arquivo texto contendo todos os nomes armazenados no SGBD. A ocorrência de um ou mais casamentos no arquivo texto fez o programa em C retornar os identificadores dos registros que casaram. Um procedimento para atualização desse arquivo texto também foi implementado e é sempre executado após uma nova carga de dados no gazetteer.

No processamento das consultas simples, inicialmente é feita uma tentativa de casamento com número de erros k=0. Se essa busca fracassa, uma nova busca é iniciada com *k* igual a 20% do comprimento da cadeia de caracteres digitada pelo usuário.

Na consulta avançada, que possui a forma <ponto de interesse> <relação espacial> <ponto de referência>, o usuário pode escolher uma das seguintes relações espaciais: “em”, “perto de”, “a 100 metros de”, “a 500 metros de”, “a 1 km de”, “a 10 km de”, “a 50 km de” e “a 100 km de”. A busca começa pela localização do ponto de referência, seguindo o mesmo procedimento descrito acima para a consulta simples. Localizado o ponto de referência, a área para a busca do ponto de interesse deve ser definida.

Se a relação “em” foi selecionada, a área de busca do ponto de interesse corresponde à própria área do ponto de referência. Para a relação “perto de”, dois casos são possíveis: se o ponto de referência é um ponto, um *buffer* de 500 metros é calculado; e se o ponto de referência é uma linha ou polígono, um *buffer* correspondente a uma expansão de 20% do retângulo envolvente mínimo do objeto é calculado. Para as demais relações, *buffers* com as distâncias selecionadas são criados.

Segue-se, então, a localização do ponto de interesse, que é limitada pela área de *buffer* calculada na etapa anterior.

4.4 Implementação do Módulo de Ranking de Lugares

Muitos lugares compartilham um nome em comum. Como exemplo, a consulta "São Francisco" retorna 683 registros no *gazetteer*: 4 municípios, 42 bairros, 629 logradouros e 8 referências. Surge desse fato a necessidade de elaborar critérios que permitam, durante uma consulta, a ordenação dos registros segundo o grau de relevância relativo dos lugares.

Para solucionar esse problema, uma heurística de atribuição de índices de importância aos lugares foi implementado. O primeiro passo foi estabelecer que o índice seria um valor real entre 0 e 1, que seria calculado sobre a população associada a cada lugar com base nos dados do IBGE. Isto significa que o registro no *gazetteer* correspondente ao Brasil possui índice de importância igual a 1 e que todos os demais registros recebem valores proporcionais a esse padrão.

Para os *territórios* cuja população é conhecida, o cálculo do índice é direto. Quando se desconhece a população, duas situações podem ocorrer. A primeira é quando se sabe a quantidade de endereços do território. Nesse caso, a população do território é estimada em 5,3 vezes o número de endereços. Este valor foi escolhido por representar o número aproximado de habitantes por endereço do município de Belo Horizonte. Quando não se sabe o número de endereços do território, a população é estimada dividindo-se a população do território superior (no caso de um bairro, o município que o contém) pela quantidade de territórios nele contidos.

Duas situações também podem ocorrer para a estimativa da população de *logradouros*. Quando se conhece o número de endereços de um logradouro, esse é multiplicado por 5,3, pelo mesmo motivo explicado anteriormente. Quando não se sabe o número de endereços, a população do logradouro é estimada como sendo a soma da população de todos os bairros que o logradouro atravessa divido pelo número total de logradouros desses bairros.

Infelizmente, o cálculo do índice de importância das *referências* não pôde ser feito seguindo a lógica utilizada para territórios e logradouros. A subjetividade de se definir a importância de um ponto de referência exigiu uma solução que começasse pelo

estabelecimento de uma classificação de 5 níveis para indicar o grau de "popularidade" de uma referência (Tabela 4.3).

Nível	Classificação	Significado
1	extremamente conhecida	conhecida em todo o país
2	muito conhecida	muito conhecida pelos moradores do município e também por pessoas de municípios vizinhos
3	conhecida	muito conhecida pelos moradores do município
4	pouco conhecida	conhecida por uma parcela dos moradores do município
5	pouquíssimo conhecida	conhecida dentro de um bairro ou pelos moradores mais próximos

Tabela 4.3 – Classificação do nível de popularidade de uma referência.

Em seguida, os índices calculados para alguns territórios foram escolhidos para que fosse possível mapear um nível da Tabela 4.3 para um valor no *ranking* de lugares. A Tabela 4.4 apresenta os municípios que foram selecionados pare esse mapeamento.

Nível	Tipo de território com popularidade semelhante	Território escolhido para o mapeamento
1	um estado	Minas Gerais
2	uma capital	Belo Horizonte
3	uma grande cidade	Juiz de Fora
4	uma cidade média	Viçosa
5	uma cidade pequena	Tiradentes

Tabela 4.4 – Mapeamento entre o nível de popularidade de uma referência e o *ranking* de lugares.

O maior inconveniente dessa abordagem é, sem dúvida, a necessidade da indicação manual do nível de popularidade das referências. Além disso, essa atribuição é subjetiva e, situação inevitável, o que pode ser considerado popular para algumas pessoas não o será para outras. Diante da impossibilidade de registrar o nível de popularidade de todas as referências no *gazetteer*, o procedimento de inserção de referências atribui o nível 3 se nenhum outro for explicitamente informado.

Processada uma consulta no Locus, a ordenação dos resultados apresentada ao usuário é elaborada com base na atribuição de um índice a cada lugar localizado. Este índice é calculado considerando-se a distância de edição calculada e o índice do lugar no *ranking* de lugares, da seguinte forma:

$$\acute{I}ndiceFinal = \acute{I}ndiceRankingDeLugares \cdot \left(1 - \left(\frac{DistânciaDeEdição}{ComprimentoDaCadeiaPesquisada}\right)\right)$$

Essa fórmula gera como resultado um índice final com valor entre 0 e 1, privilegiando valores pequenos de distância de edição e valores próximos de um no

índice de lugares (uma pequena distância de edição faz o índice final tender para um, assim como valores próximos de um no *ranking* de lugares também produzem essa tendência). A Tabela 4.5 apresenta alguns resultados da aplicação da fórmula para uma cadeia de caracteres de comprimento igual a 10.

Distância de Edição	Índice no Ranking de Lugares	Índice Final
0	1	1
1	1	0,9
1	0,9	0,81
1	0,8	0,72
2	1	0,8
2	0,9	0,72
2	0,8	0,64
3	1	0,7
3	0,9	0,63
3	0,8	0,56

Tabela 4.5 – Simulação de valores para o índice final usado na ordenação dos resultados.

4.5 Interface

Para facilitar a utilização do Locus, quatro interfaces de consulta diferentes (na verdade, páginas JSP) foram projetadas: *consulta simples*, *consulta avançada*, *consulta de endereço* e *consulta avançada com endereço*.

Na tela de *consulta simples*, o usuário digita o nome do lugar desejado e, opcionalmente, seu tipo. A Figura 4.5 apresenta um exemplo de consulta simples onde se deseja encontrar igrejas com nome São José.

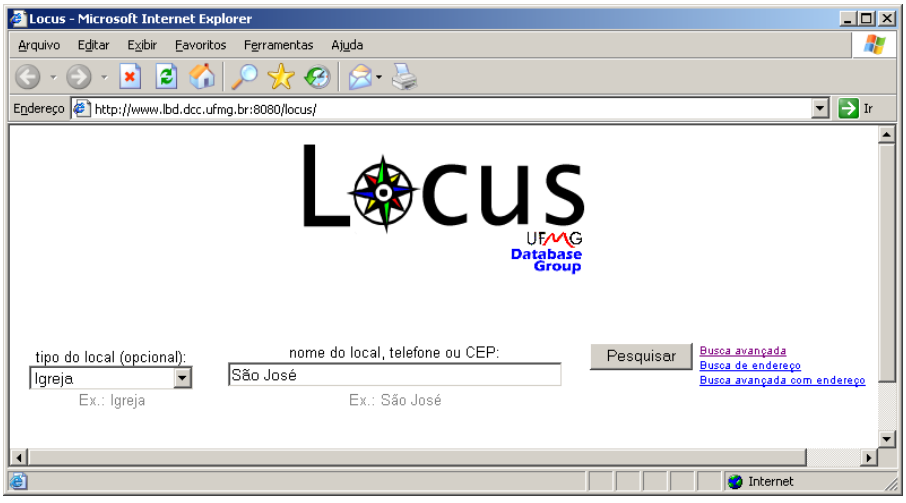


Figura 4.5 – Tela de consulta simples.

Para facilitar a seleção do usuário, os lugares encontrados são agrupados por estado (Figura 4.6). No exemplo apresentado, um único lugar foi encontrado e está localizado no bairro Centro, município de Belo Horizonte e estado de Minas Gerais.

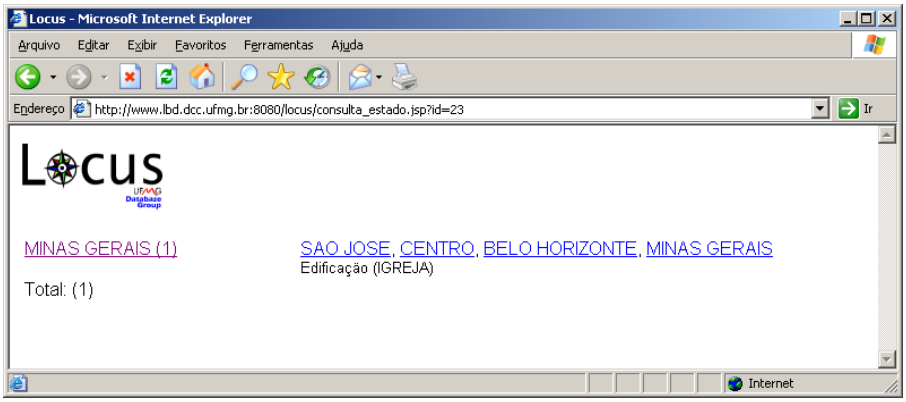


Figura 4.6 – Tela de seleção da consulta simples.

Selecionado o lugar desejado, o mesmo é apresentado na tela de resultado (Figura 4.7), que apresenta a informação de contexto, o mapa do lugar e seu MBR (retângulo envolvente mínimo).

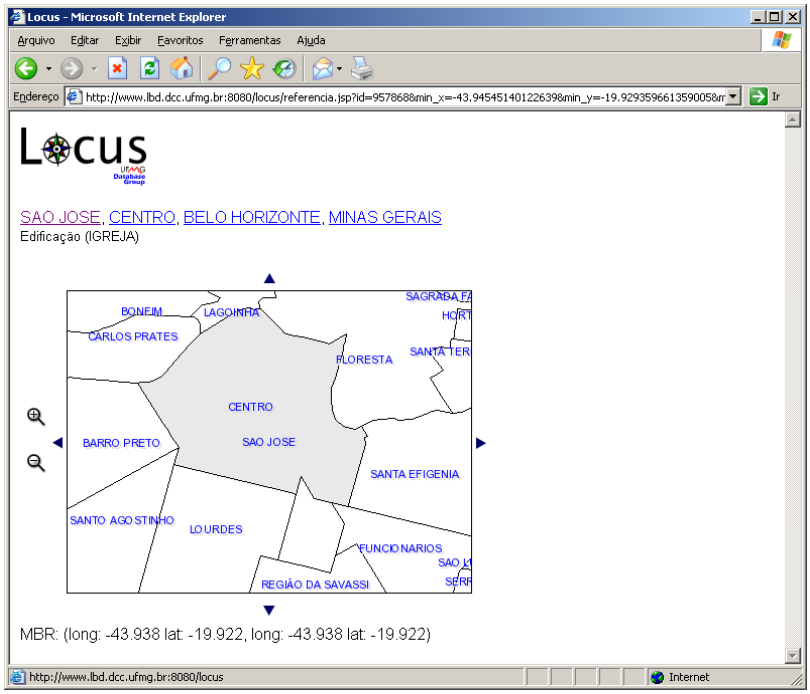


Figura 4.7 – Tela de resultado da consulta simples.

Na tela de *consulta avançada*, o usuário entra com os dados do ponto de interesse (*O que você procura*) e de referência (*Onde você procura*) desejados, além da relação espacial entre eles. A Figura 4.8 apresenta a execução da consulta "Hotel em Belo Horizonte".

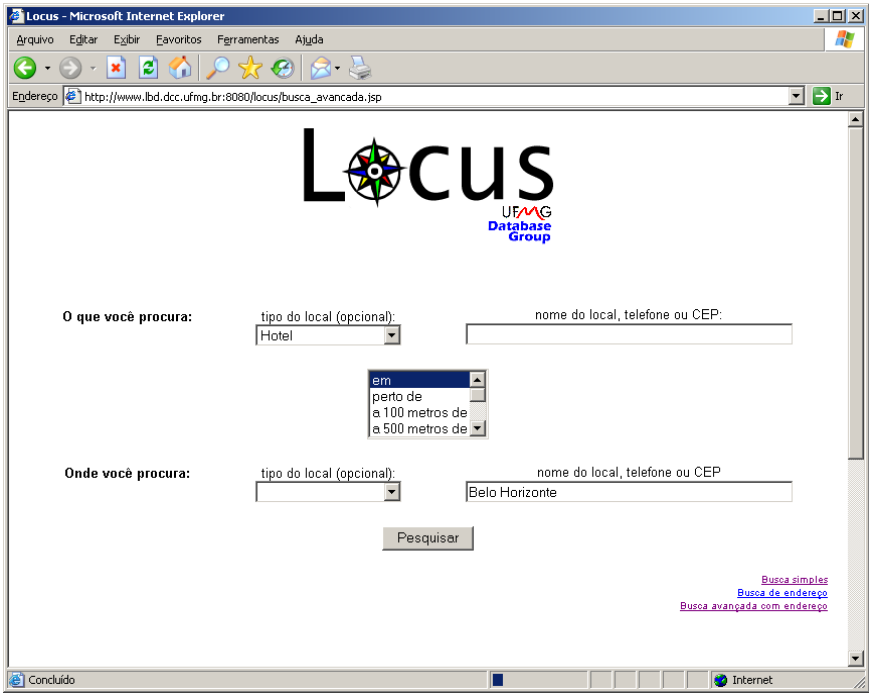


Figura 4.8 – Tela de consulta avançada.

Os pontos de referência localizados são apresentados para a seleção do usuário (Figura 4.9).

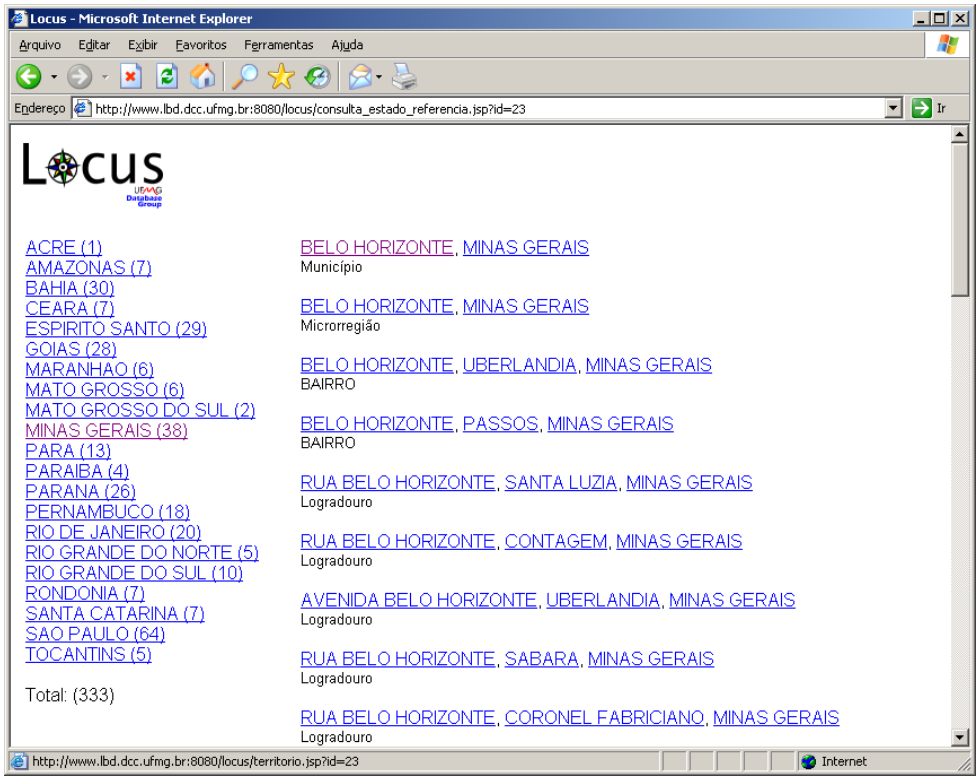


Figura 4.9 – Tela de seleção do ponto de referência na consulta avançada.

Após a seleção, as consultas espacial e textual são processadas para a localização dos pontos de interesse (Figura 4.10). A consulta termina com a seleção do lugar de interesse desejado pelo usuário.

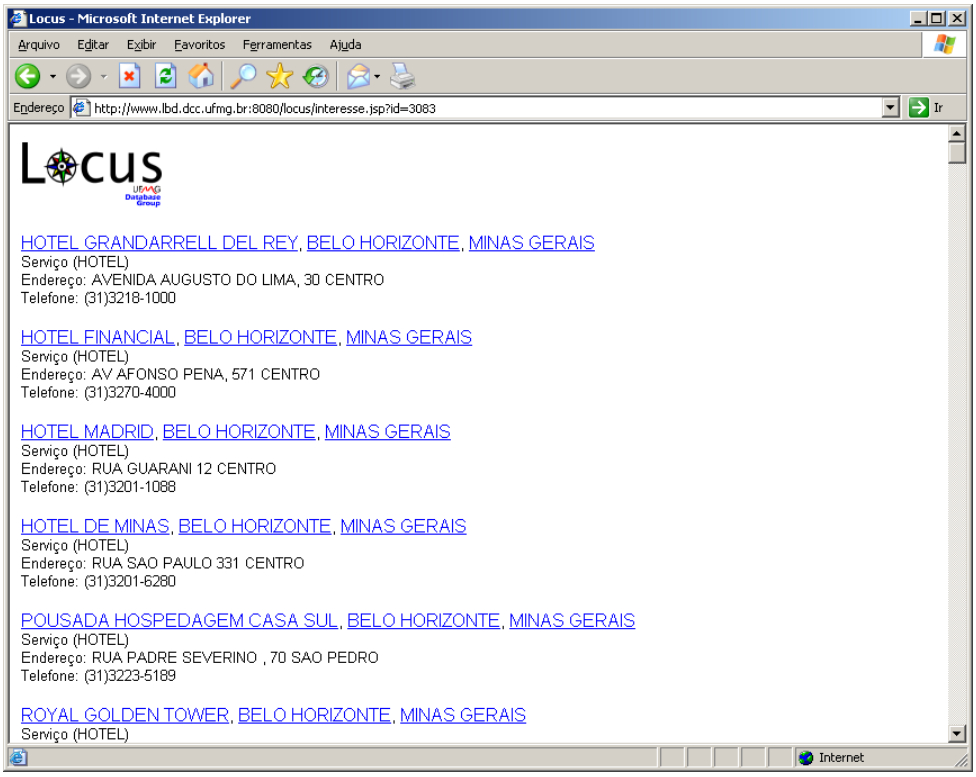


Figura 4.10 – Tela de seleção do ponto de interesse na consulta avançada.

Na *consulta por endereço*, o usuário entra com o endereço devidamente estruturado (tipo do logradouro, nome do logradouro, número, bairro e município) para que o Locus seja capaz de pesquisá-lo na base de endereços individuais armazenados no *gazetteer* (Figura 4.11).

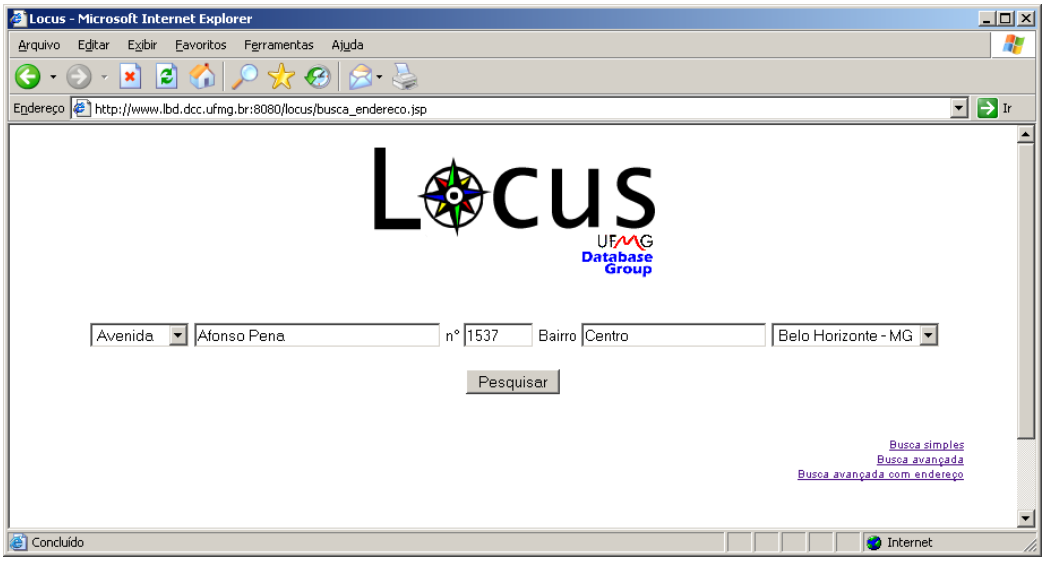


Figura 4.11 – Tela de consulta de endereço.

O resultado da consulta por endereço apresenta o endereço localizado e seus vizinhos, além da malha de *centerlines* da região (Figura 4.12).

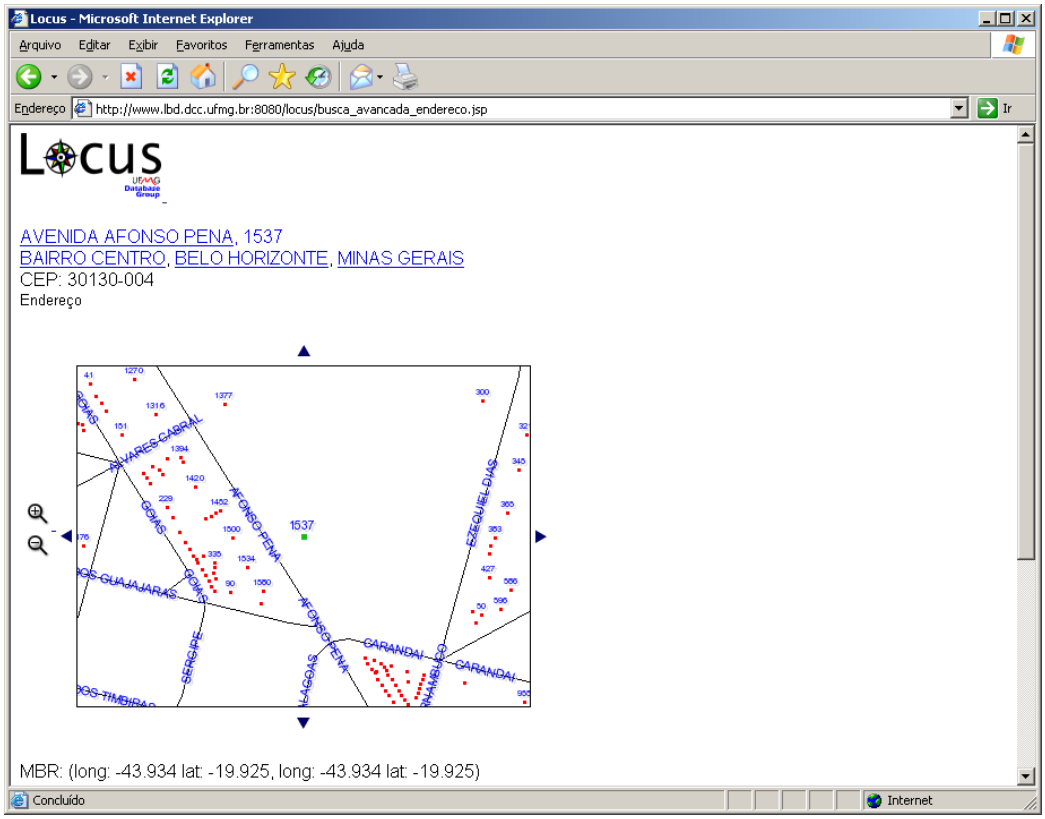


Figura 4.12 – Tela de resultado da consulta de endereço.

Por fim, na tela de *consulta avançada com endereço*, o usuário pode fazer uma consulta avançada utilizando como ponto de referência um endereço (Figura 4.13).

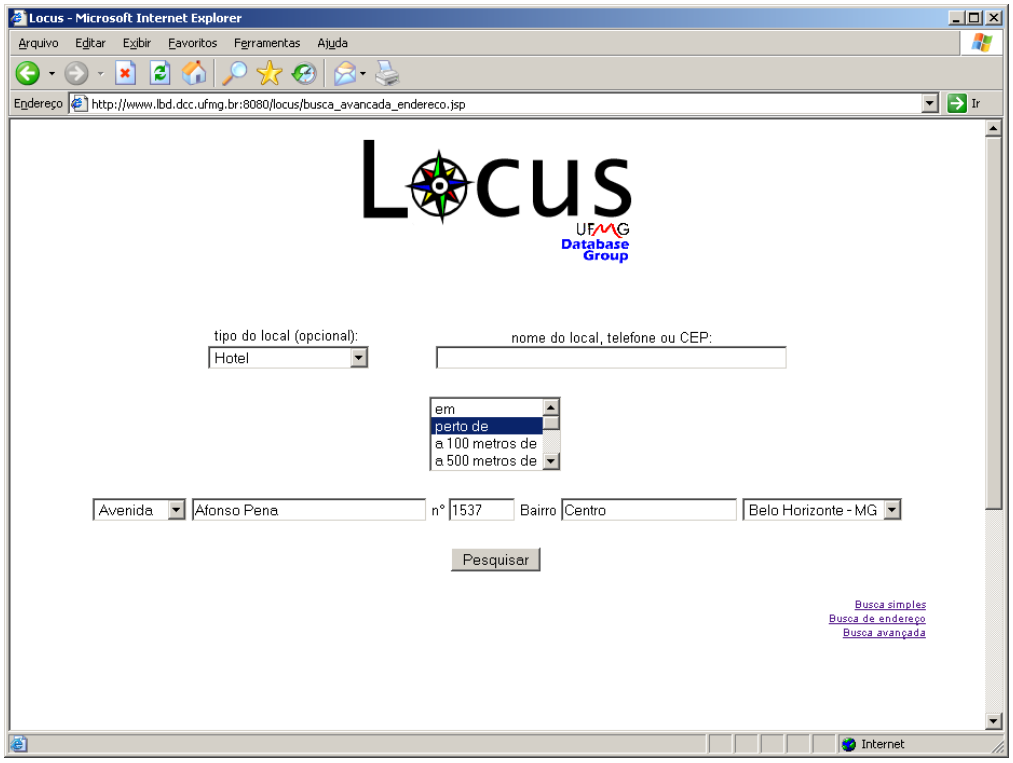


Figura 4.13 – Tela de consulta avançada com endereço.

Um exemplo de resultado da consulta é mostrado na Figura 4.14. No exemplo, são exibidos os hotéis nas proximidades do endereço informado.

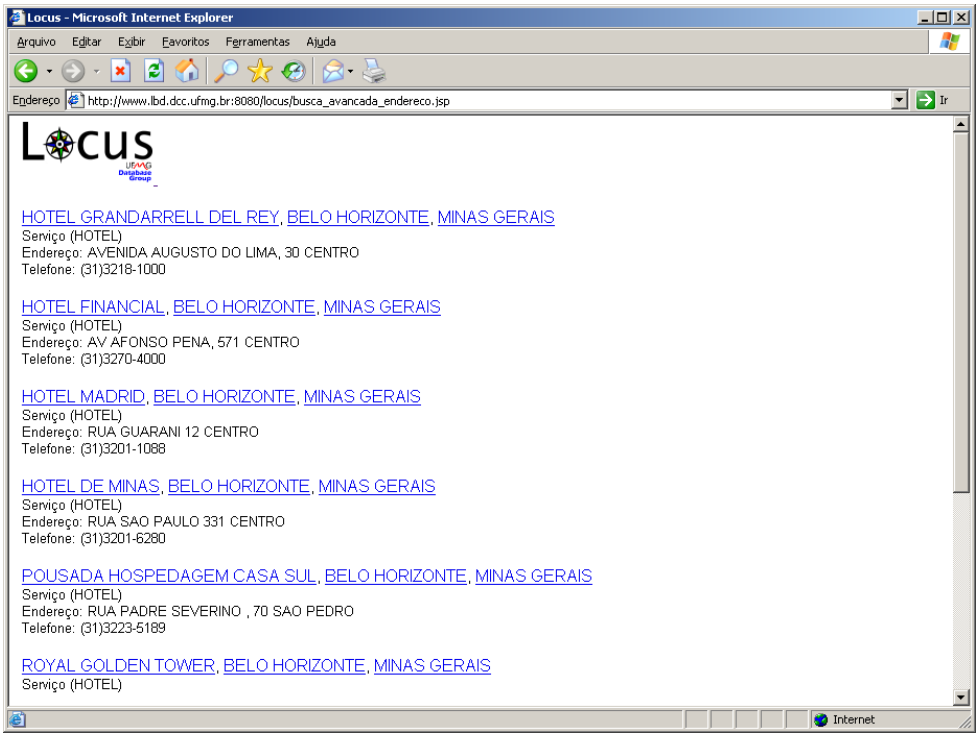


Figura 4.14 – Tela de resultado da consulta avançada com endereço.

4.6 Avaliação

Para avaliar o sistema implementado, um experimento foi conduzido com o objetivo de comparar a eficiência de uma máquina de busca atual em tratar consultas de conteúdo geográfico frente a capacidade de localização de lugares do Locus.

O experimento consistiu em aferir o sucesso da máquina de busca Google em retornar documentos relevantes para um conjunto de consultas de conteúdo geográfico e, ao mesmo tempo, aferir a capacidade do Locus em identificar os mesmos lugares presentes nas consultas avaliadas [49].

Para o experimento, inicialmente foram analisados seis meses do *log* de consultas da máquina de busca TodoBR. Foram selecionadas todas as consultas contendo algum tipo de conteúdo geográfico. Os termos utilizados para caracterizar uma consulta como sendo geográfica foram: nomes de lugar, tipos de lugar, relações espaciais e adjetivos gentílicos (mineiro, paulista, etc.). Do total de 1,4 milhão de consultas que o TodoBR recebeu nos seis meses de *log* analisados, 14,1% continham um ou mais desses termos geográficos. Esse resultado está próximo do indicado em [45].

Desse conjunto de consultas geográficas, 70 foram escolhidas aleatoriamente e distribuídas a 18 usuários, que foram instruídos a submetê-las ao Google e avaliar se a máquina de busca obteve sucesso em retornar documentos relevantes para a pesquisa. Avaliando o conjunto das 25 primeiras respostas, os usuários deveriam indicar uma das três situações: se havia documento(s) relevante(s), se havia apenas documento(s) parcialmente relevante(s) ou se não havia nenhum documento relevante. O critério de classificação da qualidade de resposta em uma dessas três categorias baseou-se puramente no bom senso dos usuários, já que não lhes foi passada nenhuma regra sobre como deveriam determinar a relevância dos documentos. O conteúdo do questionário enviado aos usuários aparece na Figura 4.15.

A Figura 4.16 apresenta o gráfico do resultado dessa parte do experimento. A máquina de busca foi capaz de retornar documentos relevantes para 45 das 70 consultas (64,3%). Os usuários ficaram parcialmente satisfeitos para 13 consultas (18,6%) e absolutamente não satisfeitos para 12 consultas (17,1%).

Questionário de Avaliação

Para cada consulta que você recebeu, por favor responda:

1) A consulta é espacial (sim ou não)?

No seu entendimento, a consulta possui um ou mais termos geográficos, como nomes de lugares (ex.: Belo Horizonte, Rio de Janeiro). Se a resposta for "não", passe para a próxima consulta.

2) Qual a qualidade da resposta do Google (atende completamente, atende parcialmente ou não atende)?

Digite a consulta, SEM MODIFICÁ-LA, no Google. Considerando as 25 primeiras páginas retornadas na consulta, como você avalia a qualidade da resposta? Em outras palavras, alguma das páginas retornadas possui conteúdo que corresponde ao que, no seu entendimento, o usuário procura?

3) Você encontrou o lugar no Locus (sim ou não)?

Tente encontrar no Locus o lugar indicado na consulta. Nesse caso, NÃO DIGITE A CONSULTA COMPLETA no Locus, mas somente aquilo que você identificou como sendo um lugar. Por exemplo: em "juizado em Natal", tente encontrar Natal; em "bar na Savassi", tente encontrar Savassi. Você pode especificar o tipo do lugar procurado ou usar a consulta avançada para localizar o lugar. Da mesma forma como no Google, considere apenas os 25 primeiros lugares recuperados pelo Locus.

Figura 4.15 – Questionário de avaliação do experimento.

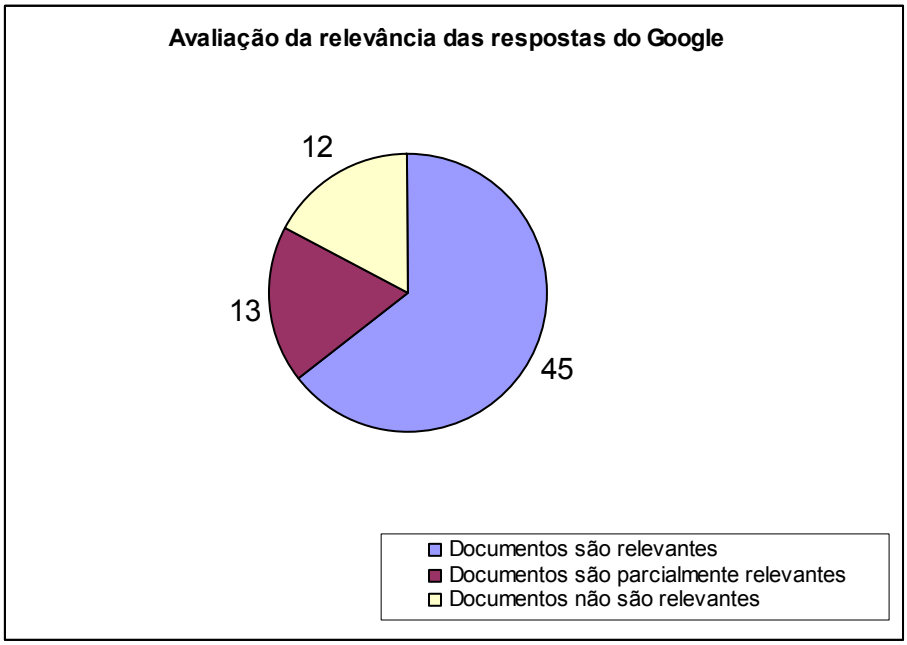


Figura 4.16 – Avaliação da relevância das respostas do Google.

Na seqüência, os usuários foram instruídos a tentar encontrar no Locus os mesmos lugares indicados nas consultas que foram submetidas ao Google. Os usuários ficaram livres para utilizarem tanto a busca simples ou avançada do Locus. Para tornar a comparação justa, também foram consideradas apenas as 25 primeiras respostas do Locus. Segundo os usuários, o Locus obteve êxito em encontrar o lugar correto para 54 das 70 consultas (Figura 4.17), o que corresponde a uma taxa de sucesso de 77,1%.

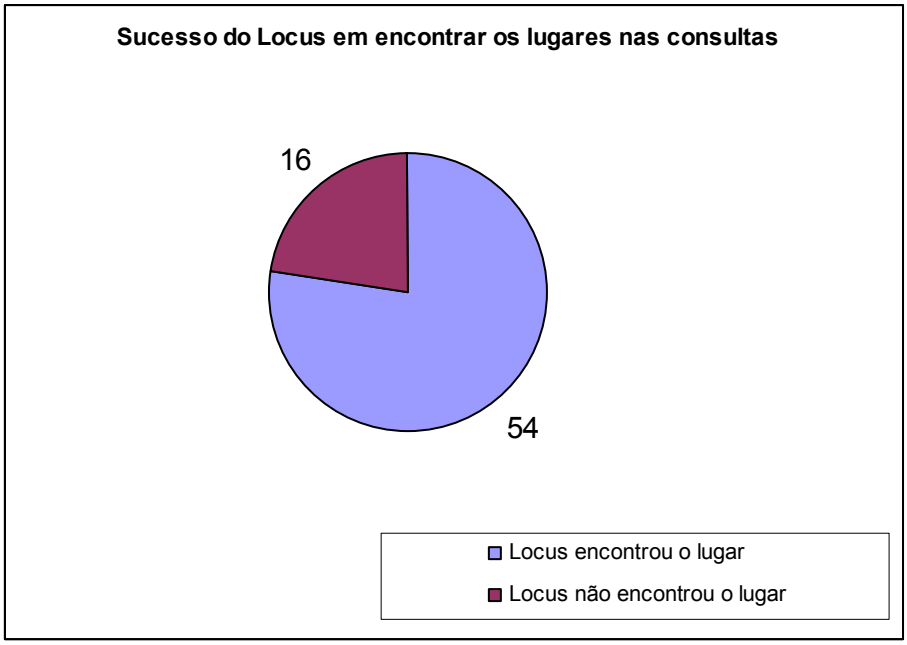


Figura 4.17 – Sucesso do Locus em encontrar os lugares indicados nas consultas.

As consultas para as quais o Locus falhou em encontrar o lugar representaram, em sua maioria, dois tipos de situação: o lugar em questão era uma referência estrangeira ou o lugar era uma referência brasileira que, de fato, não estava cadastrada no gazetteer.

O resultado da avaliação foi considerado positivo, já que o Locus obteve mais sucesso em encontrar os lugares do que a máquina de busca em retornar documentos relevantes para a consulta. Como a máquina de busca está ignorando a semântica espacial durante a indexação dos documentos, muitos resultados relevantes não aparecem nas primeiras posições do *ranking*. Nossa convicção é de que futuras melhorias no desempenho de máquinas de busca para o caso de buscas com conteúdo geográfico dependerão do uso do conhecimento geográfico contido em *gazetteers* nas fases de indexação dos documentos e de processamento das consultas.

5 Conclusões e Trabalhos Futuros

Este trabalho apresentou o Locus, um sistema de localização através de referências espaciais indiretas. O sistema fundamenta-se em uma ontologia de lugar. Essa ontologia modela os conceitos do espaço geográfico, especialmente do espaço urbano. A ontologia de lugar já se encontrava especificada antes da elaboração e construção do Locus. Ela não é, portanto, uma contribuição deste trabalho. Nossa contribuição principal consistiu em especificar e implementar um sistema que refletisse o conhecimento representado nessa ontologia.

Da ontologia, foi especificado o esquema conceitual do *gazetteer*, que representa diversas entidades presentes no espaço geográfico urbano, como pontos de referência, endereços, CEPs e números de telefone.

Os dados carregados foram fornecidos por diversos órgãos públicos e também extraídos da Web, com as ferramentas ASByE e DEByE. A extração de referências espaciais da Web, como descrito na Seção 4.2, demonstrou ser um procedimento viável e muito eficiente.

A busca foi implementada de forma a admitir erros nos dados digitados pelo usuário A consulta pode ser feita utilizando relações espaciais no formato "ponto de interesse", "relação espacial", "ponto de referência". Os resultados retornados são ordenados de acordo com o grau de importância relativo dos lugares.

O sistema implementado é totalmente baseado em software livre e utiliza o PostGIS, um SGBD com capacidade de armazenamento de dados espaciais. É possível utilizar o Locus de duas formas: por meio de uma interface Web ou via serviço Web (WFS – Web Feature Service).

A abordagem seguida fez com que o sistema apresentasse diferenciais importantes em relação a outros *gazetteers* atualmente disponíveis. Entre eles:

- O processamento de consultas espaciais: o Locus admite consultas no formato <ponto de interesse> <relação espacial> <ponto de referência>.
- O tratamento de referências intra-urbanas, incluindo endereços individuais: os gazetteers atuais não armazenam esse tipo de referência, especialmente serviços.

- A possibilidade de expansão semi-automática do *gazetteer*, com a alimentação de referências extraídas da Web.
- A navegação guiada pelo contexto geográfico das referências: as telas da interface Web do Locus são formatadas de acordo com o tipo do lugar localizado e apresentam opções de consulta e navegação a outros lugares relacionados.
- O armazenamento de referências não limitado apenas a nomes de lugares, mas incluindo também CEPs e números de telefone.
- O processamento de consultas admitindo erros nos nomes das referências: os *gazetteers* atuais admitem apenas busca exata ou, no máximo, busca por *substring* (o nome informado na pesquisa deve estar contido no nome do lugar), o que torna a busca limitada.

O resultado final é um sistema com potencial de utilização em várias áreas de aplicação onde o reconhecimento de referências geográficas seja necessário. Dado o interesse crescente na área de recuperação de informação geográfica, acreditamos que o Locus possa atender satisfatoriamente uma grande quantidade de novos sistemas.

Alguns trabalhos atualmente em desenvolvimento no Laboratório de Bancos de Dados da UFMG utilizam o Locus. Um deles estuda o reconhecimento de evidências geográficas em páginas Web, o que necessariamente depende de uma base de conhecimento geográfico tal qual a que o Locus é capaz de fornecer. Outro se concentra na interpretação das expressões geográficas que utilizamos cotidianamente para nos localizarmos, como "hotel *perto da* Praça da Liberdade" ou "posto de gasolina *a 1.000 metros* do BH Shopping" [10]. O Locus tem sido utilizado nos dois projetos e tem demonstrado ser capaz de atender suas necessidades. Um terceiro trabalho utiliza o Locus para a busca georreferenciada de objetos em uma biblioteca digital de dados ecológicos, a BDiGPELD - Biblioteca Digital Georreferenciada para Pesquisas Ecológicas de Longa Duração [3].

Há, também, diversas aplicações para o Locus no campo do georreferenciamento. O custo de coleta de dados espaciais é alto e, para a grande maioria das aplicações, não se exige grande precisão nos dados. O georreferenciamento automático é uma solução de baixo custo e adequada para muitos casos. O Locus é capaz de dar suporte a sistemas de georreferenciamento automático, incluindo a geocodificação aproximada de endereços

[9]. São muitas as áreas onde essa solução é interessante, entre elas a saúde pública, educação e segurança.

Acreditamos ser viável, ainda, a utilização do Locus em sistemas de navegação pessoal, como os que já começam a ser incorporados aos telefones móveis mais modernos. Assim como também em aplicações geográficas interativas na Web e quaisquer outros sistemas baseados em localização.

Como sugestões para trabalhos futuros, pode-se indicar:

- o estudo e elaboração de um melhor *ranking* de entidades geográficas. O *ranking* implementado no Locus é limitado, especialmente no que diz respeito às referências (serviços, acidentes geográficos, etc.). Uma sugestão seria utilizar a freqüência com a qual uma entidade é referenciada em páginas da Web. Há aí, entretanto, a dificuldade relacionada à ambigüidade semântica sempre presente em nomes de lugares. Parece claro que um *ranking* apenas será satisfatório se considerar aspectos da intenção do usuário, ou do contexto em que a busca está sendo conduzida;
- um experimento semelhante ao apresentado no Capítulo 5, mas com um número maior de usuários e incluindo uma avaliação do perfil desses usuários;
- um experimento que compare os resultados do ranking de lugares implementado no Locus frente a uma ordenação processada com base no modelo vetorial;
- a implementação de um serviço Web mais completo. O Locus pode ser acessado via WFS – *Web Feature Service*. Este serviço, porém, possui opções básicas de recuperação de entidades geográficas, que não incluem a possibilidade de busca admitindo erro ou o processamento de consultas espaciais.

6 Referências bibliográficas

[1] ADL - Alexandria Digital Library Gazetteer. Disponível em: <http://www.alexandria.ucsb.edu>. Acesso em: 08 maio 2005.

[2] BARROS, E. G., LAENDER, A. H. F. Uma biblioteca digital para o PELD Brasil. **Resumos do Simpósio Internacional sobre Projetos Ecológicos de Longa Duração**, Manaus, pp. 57-59, 2004.

[3] BARROS, E. G., SOUZA, L. A., COTA, R. G., LAENDER, A. H. F., BORGES K. A. V., GONÇALVES, M. Uma Biblioteca Digital Georreferenciada para Dados Ecológicos. **Simpósio Brasileiro de Banco de Dados**, Uberlândia, 2005.

[4] BERNERS-LEE, J., HENDLER, J., LASSILA, O. The Semantic Web. **Scientific American**, v. 184, n. 5, pp. 34-43, 2001.

[5] BOOCH, G., JACOBSON, I., RUMBAUGH, J. **The Unified Modeling Language User Guide**. Addison-Wesley, Reading – MA, 1999.

[6] BORGES, K. A. V., DAVIS JR., C. A., LAENDER, A. H. F. OMT-G: An object-oriented data model for geographic applications. **GeoInformatica**., v. 5, n. 3, pp. 221-260, 2001.

[7] BORGES, K. A. V., DELBONI, T. M., SOUZA, L. A., LAENDER, A. H. F., DAVIS Jr., C. D. A. Determination of Approximate Locations from Web Pages Based on an Ontology of Place and Spatial Relations. Em preparação.

[8] BORGES, K. A. V., LAENDER, A. H. F., MEDEIROS, C. B., SILVA, A. S., DAVIS JR., C. A. The Web as a data source for spatial databases. **V Geoinfo – Simpósio Brasileiro de Geoinformática**, Campos do Jordão, 2003. CD-ROM.

[9] DAVIS JR., C. A., FONSECA, F. T., BORGES, K. A. V. A Flexible Addressing System for Approximate Geocoding. **V Geoinfo – Simpósio Brasileiro de Geoinformática**, Campos do Jordão, 2003. CD-ROM.

[10] DELBONI, T. M., Expressões de Posicionamento como Fonte de Contexto Geográfico na Web, Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

[11] EGENHOFER, M. J. Toward the semantic geospatial web. **Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems**, McLean, Virginia, pp. 1-4, 2002.

[12] EGENHOFER, M. J., FRANZOSA, R. D. Point-set topological spatial relations. **International Journal of Geographical Information Systems**, v. 5,

n. 2, pp. 161-174, 1991.

[13] FARRELL, C., SCHULZE, M., PLEITNER, S., DALBONI, D. DNS Encoding of Geographical Location. Disponível em <http://www.apps.ietf.org/rfc/ rfc1712.html>. Acesso em: 21 março 2005.

[14] FINCH, D. Specification of System Functionality. Disponível em <http://www.geo-spirit.org/publications/ SPIRIT_WP1_D4.pdf>. Acesso em 15 março 2005.

[15] FONSECA, F. T., DAVIS, C. A., CÂMARA, G. Bridging Ontologies and Conceptual Schemas in Geographic Information Integration. **GeoInformatica**, v. 7, n. 4, pp. 355-378, 2003.

[16] FU, G., ABDELMOTY, A. I., JONES A. B. Design of a Geographical Ontology. Disponível em <http://www.geo-spirit.org/publications/ SPIRIT_WP3_D5.pdf>. Acesso em 15 março 2005.

[17] GOLGHER, P. B., LAENDER, A. H. F., SILVA, A. S., RIBEIRO-NETO, B. A. An Example-Based Environment for Wrapper Generation. **Proceedings of the Workshops on Conceptual Modeling Approaches for E-Business and The World Wide Web and Conceptual Modeling: Conceptual Modeling for E-Business and the Web**, Salt Lake City, Utah, Lecture Notes in Computer Science, v. 1921, pp. 152-164, 2000,

[18] GRUBER, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal of Human and Computer Studies**, v. 43, n. 5/6, pp. 907-928, 1995.

[19] GUARINO, N. Formal Ontologies and Information Systems, **Formal Ontologies and Information Systems**, Ed. Amsterdan, Netherlands: IOS Press, pp. 3-15, 1998.

[20] GUTTMAN, A. R-trees: a dynamic index structure for spatial searching. **Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data**, Boston, Massachusetts, pp. 47-57, 1984.

[21] HILL, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. **Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries**, Lisboa, Portugal, pp. 280-290, 2000.

[22] HILL, L. L. Workshop Report on Digital Gazetteers: Integration into Distributed Digital Library Services, **Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital libraries**, Portland, Oregon, pp. 427, 2002.

[23] HILL, L. L., GOODCHILD, M. F. Digital Gazetteer Information Exchange Workshop Final Report. Disponível em <www.alexandria.ucsb.edu/gazetteer/dgie/DGIE_website/DGIEfinal.pdf>. Acesso em 28 março 2005.

[24] ICANN - The Internet Corporation for Assigned Names and Numbers. **Summary Application of SRI International**. Disponível em <http://www.icann.org/tlds/report/geo1.html>. Acesso em 09 maio 2005.

[25] JANEÉ, G., FREW, J., HILL, L. L., Issues in Georeferenced Digital Libraries. **D-Lib Magazine**, v. 10, n. 5, 2004.

[26] JONES, C. B , ALANI, H., TUDHOPE, D. Geographical information retrieval with ontologies of place. **Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science**, Lecture Notes in Computer Science, v. 2205, p.322-335, 2001.

[27] JONES, C. B., Purves, R., Ruas, A, Sanderson, M., Sester, M., van Kreveld, M., Weibel, R. Spatial Information Retrieval and Geographical Ontologies. An Overview of the SPIRIT Project. **Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, Tampere, Finland, pp. 387-388, 2002.

[28] LAENDER, A. H. F., BORGES, K. A. V., CARVALHO, J. C. P., MEDEIROS, C. B., SILVA, A. S., DAVIS JR., C. A. . Integrating Web Data and Geographic Knowledge Into Spatial Databases. **In: Yannis Manolopoulos; Apostolos Papadopoulos; Michael Vassilakopoulos. (Org.). Spatial Databases: Technologies, Techniques and Trends.** Hershey, Pennsylvania, pp. 23-48, 2004.

[29] LAENDER, A. H. F., RIBEIRO-NETO, B. A., SILVA, A. S., TEIXEIRA, J. S. A Brief Survey of Web Data Extraction Tools. **ACM SIGMOD Record**, v. 31, n. 2, 2002.

[30] LAENDER, A. H. F., SILVA, A. S., GOLGHER, P. B., RIBEIRO-NETO, B. A., EVANGELISTA-FILHA, I. M. R., MAGALHÃES, K. V. The DEByE Environment for Web Data Management. **IEEE Internet Computing**, v. 6, n. 4, 2002.

[31] LARSON, R. R. Geographic Information Retrieval and Spatial Browsing. Disponível em < http://sherlock.berkeley.edu/geo_ir/part1.html>. Acesso em 28 março 2005.

[32] LEIDNER, J. L. Toponym Resolution in Text: "Which Sheffield is it?", **Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieva**l, Sheffield, pp. 602, 2004.

[33] MCCURLEY, K. S. Geospatial Mapping and Navigation of the Web. **International World Wide Web Conference**, Hong Kong, pp. 221-229, 2001.

[34] NAVARRO, G. A. Guided Tour to Approximated String Matching. **ACM Computing Surveys**, v. 33, n. 1, pp. 31-88, 2001.

[35] OLLIGSCHLAEGER. A. M., HAUPTMANN, A. G. Multimodal Information Systems and GIS: The Informedia Digital Video Library. **Proceedings of the ESRI User Conference**, San Diego - CA, 1999.

[36] OpenGIS Gazetteer Service Profile of a WFS. Disponível em <http://www.opengeospatial.org/specs>. Acesso em 28 março 2005.

[37] OpenGIS Geocoder Service Specification - Discussion Paper. Disponível em <http://www.opengeospatial.org/specs>. Acesso em 28 março 2005.

[38] OpenGIS Geography Markup Language - Implementation Specification. Disponível em <http://www.opengeospatial.org/specs>. Acesso em 28 março 2005.

[39] OpenGIS Geoparser Service Specification - Discussion Paper. Disponível em <http://www.opengeospatial.org/specs>. Acesso em 28 março 2005.

[40] OpenGIS Simple Features Specification for SQL - Implementation Specification. Disponível em <http://www.opengeospatial.org/specs>. Acesso em 28 março 2005.

[41] OpenGIS Web Feature Service - Implementation Specification. Disponível em <http://www.opengeospatial.org/specs>. Acesso em 28 março 2005.

[42] PINTO, M. V. Cadastramento escolar: democratização do acesso à escola pública. **Informática Pública**, v. 1, n. 2, pp. 139-156, 1999.

[43] PURVES, R., JONES, C. Workshop on geographic information retrieval. **ACM SIGIR Forum**, v. 38, n. 2., pp. 53-56, 2004.

[44] SAMET, H., The Quadtree and Related Hierarchical Data Structures. **ACM Computing Surveys**, v. 16, n. 2, pp. 187-260, 1984.

[45] SANDERSON, M., KOHLER, J. Analyzing Geographic Queries. **ACM SIGIR 2004 - Workshop on Geographic Information Retrieval**. Disponível em <http://www.geo.unizh.ch/~rsp/gir/abstracts/sanderson.pdf>. Acesso em 08 maio 2005.

[46] SAUDAVEL - Sistema de Apoio Unificado para Detecção e Acompanhamento em Vigilância EpidemioLógica. Disponível em <http://saudavel.dpi.inpe.br>. Acesso em 08 maio 2005.

[47] SILVA, M. J. Adding Geographic Scopes to Web Resources. **ACM SIGIR 2004 - Workshop on Geographic Information Retrieval**. Disponível em <http://www.geo.unizh.ch/~rsp/gir/abstracts/silvia.pdf>. Acesso em 08 maio 2005.

[48] SMITH, D. A., MANN, G. S. Bootstrapping Toponym Classifiers. **HLT-NAACL Workshop: Analysis of Geographic References**, Edmonton – Alberta, pp. 45-49, 2003.

[49] SOUZA, L. A., DAVIS JR., C. A., BORGES, K. A. V., DELBONI, T. M., LAENDER A. H. F. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. 3th LA-Web - Latin American Web Congress, Buenos Aires, 2005.

[50] SOUZA, L. A., DELBONI, T. M., BORGES K. A. V., DAVIS JR., C. A., LAENDER A. H. F. Locus: Um Localizador Espacial Urbano, **VI Geoinfo – Simpósio Brasileiro de Geoinformática**, Campos do Jordão, pp. 467-478, 2004.

[51] WOODRUFF, A. G., PLAUNT, C. GIPSY: Geo-referenced Information Processing System. **Journal of the American Society for Information Science**, v. 45, pp. 645-655, 1994.

[52] WU, S., MANBER, U. Fast Text Searching Allowing Errors. **Communications of The ACM**, v. 35, n. 10, pp. 83-91, 1992.

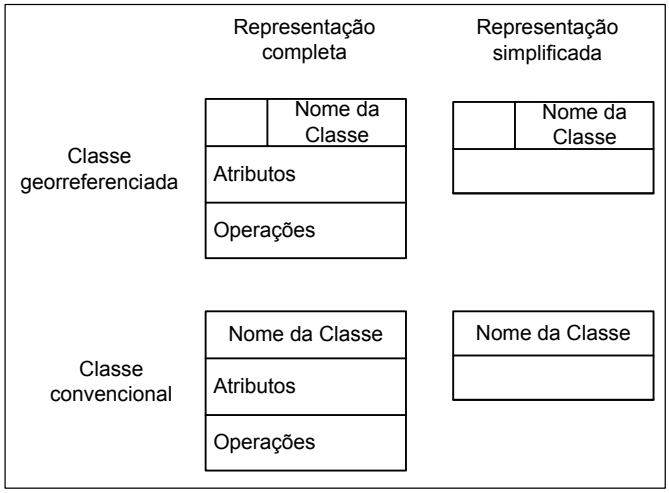
[53] ZIVIANI, N. **Projeto de Algoritmos com Implementações em Pascal e C – Segunda Edição**. Pioneira Thompson Learning, São Paulo, 2004.

[54] ZONG, W., WU, D., SUN, A., LIM, E. P., GOH, D. H., On Assigning Place Names to Geographic Related Web Pages. **Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries**, Denver, CO, pp. 354-362, 2005.

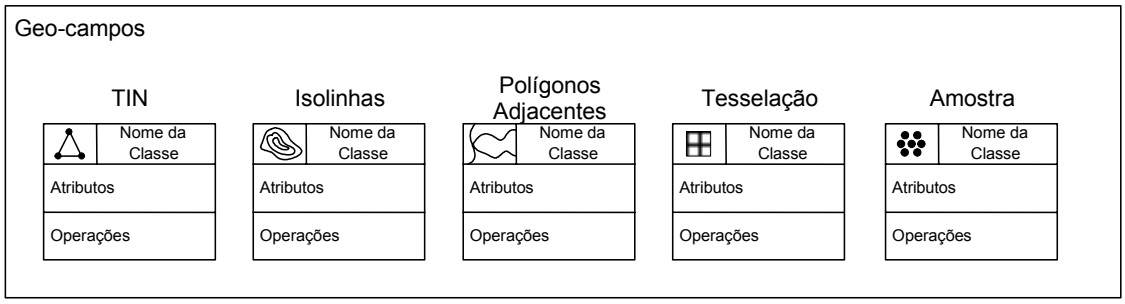
Apêndice – Notação OMT–G

O esquema conceitual do Locus foi especificado com o modelo para aplicações geográficas OMT-G. O modelo utiliza construtores da linguagem UML e inclui primitivas específicas para representar a geometria e a topologia de dados geográficos.

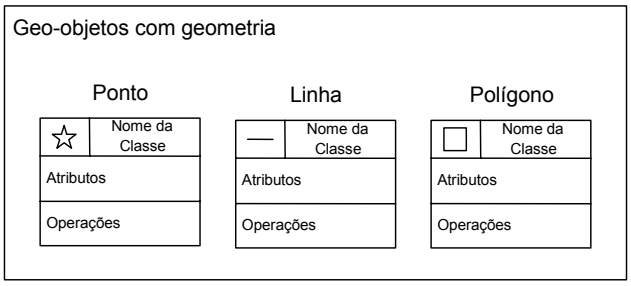
Três conceitos formam a base do modelo: *classes*, *relacionamentos* e *restrições de integridade*. Classes e relacionamentos são as duas primitivas básicas para criar os esquemas das aplicações. Um diagrama de classes OMT-G define a estrutura e o conteúdo de um banco de dados geográfico. As classes podem ser georreferenciadas ou convencionais:



As classes georreferenciadas especializam-se em geo-campos (para representar objetos e fenômenos com distribuição contínua sobre o espaço) ou geo-objetos (para representar objetos geográficos discretos). Os geo-campos especializam-se nas classes TIN, Isolinhas, Polígonos Adjacentes, Tesselação e Amostra:



Os geo-objetos especializam-se em geo-objetos com geometria e geo-objetos com geometria e topologia. Os geo-objetos com geometria são Ponto, Linha e Polígono:



Os geo-objetos com geometria e topologia são Arco Unidirecional, Arco Bidirecional e Nó: