Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Web

Karla Albuquerque de Vasconcelos Borges

Belo Horizonte, Minas Gerais
2006

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Web

Karla Albuquerque de Vasconcelos Borges

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para obtenção do grau de Doutor em Ciência da Computação

Orientador: Alberto Henrique Frade Laender
Co-orientadora: Claudia Maria Bauzer Medeiros

Belo Horizonte, Minas Gerais
2006

Agradecimentos

Agradeço a toda minha família pelo apoio e incentivo que sempre me deram no decorrer dessa longa caminhada. Em especial, ao Romero, que esteve mais uma vez enfretando o desafio de ser marido de pós-graduanda. Agradeço o seu imenso amor e apoio, sem os quais teria sido muito difícil chegar ao fim. Também muito especial é o agradecimento ao meu filho Raphael que sempre foi capaz de me apoiar e entender as várias vezes em que eu não podia ficar muito tempo por perto. O seu carinho e apoio me deram um banho de ânimo nos vários momentos de fraqueza.

À minha mãe e ao meu pai que, sempre presentes com seu amor e carinho, me ensinaram a ser persistente e a enfrentar os desafios da vida.

Aos meus queridos orientadores, Alberto e Claudia pela orientação tão criteriosa responsável pela qualidade deste trabalho. Agradeço também pelo carinho e atenção sempre a mim dispensados e, principalmente, pelo apoio e compreensão diante dos momentos difíceis que passei com a perda do meu pai.

Às meninas da secretaria do DCC, sempre tão eficientes, por toda ajuda e, principalmente, pela amizade que surgiu entre nós. À Renata, Túlia, Sheila e Maristela, o meu muito obrigada.

À vida por colocar na minha trajetória amigos tão queridos, que não só compartilharam momentos difíceis e felizes, como também fizeram com que esta caminhada fosse um dos melhores momentos da minha vida. À Joyce, Kissia, Lili, Silvana, Andréa, Umberto, Olga, Delboni, Lena, Patrícia, Guilherme, Allan, Ruiter e Evandrino, o meu eterno agradecimento. Aos amigos Pio e Ligiane, um agradecimento especial.

Aos amigos do LBD pela convivência, pela troca de experiências e pela amizade. Em especial agradeço ao Altigran, Ligiane e ao Delboni por entender e compartilhar minhas idéias fazendo com que minha caminhada não fosse solitária. Agradeço pelas discussões proveitosas, pelas sugestões e por toda ajuda técnica.

Aos amigos da Prodabel em especial ao Marco Antônio por toda ajuda. À Dilu, Sandra, Fatinha, Márcio, Angelo, Márcio Venuto, Pedro, Glauciene, Thelma, Myrza e Marquinhos pelo carinho e incentivo. À Lili, Piedade e Taciana pelo companheirismo.

À Sivakami e Renata, pelos ensinamentos de vida que me mantiveram firme no decorrer do doutorado.

Aos meus amigos Fred e Clodoveu pela presença amiga constante e pelas discussões que tanto me ajudaram.

À Diretoria da Prodabel pela oportunidade concedida através do seu programa de doutorado.

Finalmente, agradeço ao Espírito Infinito, que sempre guia e ilumina os meus caminhos, pela vitória alcançada.

Om Gum Gurave Namah

Ao meu pai Albino

Resumo

Consultas que incluem pelo menos um termo relacionado a geografia, como nomes de lugar e feições naturais, são hoje um subconjunto significativo das consultas submetidas às máquinas de busca. O interesse por informação local na Web (busca local) vem aumentando a cada dia e para esse tipo de busca, a Web é um vasto repositório de informação local e geográfica. No entanto, as máquinas de busca tradicionais apresentam limitações quanto ao reconhecimento do escopo geográfico existente nas páginas da Web. Páginas referentes ao mesmo lugar, mas que usam nomes alternativos provavelmente não serão recuperadas juntas. Além disso, muitas vezes o contexto geográfico existente nas páginas está implícito, podendo ser inferido pela existência, por exemplo, de um número de telefone ou código postal.

Para resolver esses problemas, esta tese tem como foco a *Web local*, propondo uma abordagem apoiada em uma ontologia de lugar urbano, que permita reconhecer, extrair e geocodificar evidências geo-espaciais de características locais, como endereços, códigos postais e telefones presentes em páginas da Web. As evidências geo-espaciais representam localizações implícitas, capazes de correlacionar o conteúdo de uma página, ou de parte dela, a uma localização geográfica urbana. Assim, as máquinas de busca poderiam por exemplo, utilizar essa informação para a recuperação de páginas referentes a serviços e atividades em uma determinada localidade ou próximos a ela.

Assim as principais contribuições desta tese são (1) caracterização de endereços presentes em páginas da Web como fontes de evidência geo-espacial e definição de padrões para o seu reconhecimento e extração, (2) definição da *OnLocus*, uma ontologia de lugar urbano para auxiliar o processo de reconhecimento e extração de evidências geo-espaciais de páginas da Web, (3) criação de uma base de conhecimento para reconhecimento de lugares brasileiros, baseada na *OnLocus*, (4) proposta de uma estratégia de categorização geográfica de uma página,ou de partes dela, dentro da divisão territorial de um país, e (5) avaliação das características quantitativas e qualitativas dos endereços presentes nas páginas da Web brasileira. Todas essas contribuições foram validadas por meio de experimentação, usando dados reais correspondentes a um conjunto de 4 milhões de páginas da Web. Como conseqüência

adicional, foi possível traçar um retrato das páginas da Web brasileira no que tange a padrões de endereço e, conseqüentemente, entender melhor como geocodificá-las. Os resultados desta tese abrem um leque de perspectivas para novos tipos de aplicação, como, por exemplo, uso de *links* de navegação baseados em localização geográfica, classificação geográfica das páginas da Web, mineração de dados geo-espaciais em páginas da Web e anotação semântica das páginas.

Abstract

Queries that include at least one geographic-related term, such as place names and natural features, are currently a significant subset of the queries that are submitted to search engines. Interest on local information on the Web (local search) is increasing daily, and for this kind of search, the Web is a vast repository of local geographic information. However, traditional search engines have limitation on the recognition of the geographic scope of Web pages. Pages that refer to the same place, but using alternative names, probably will not be retrieved together. Besides, in many situations the geographic context is implicit in the pages, but can be inferred by the existence, for instance, of a telephone number or postal code.

In order to propose a solution for these problems, this thesis focuses on the *local Web*, presenting an approach based on an ontology of urban place, which allows for the recognition, extraction, and geocoding of geospatial evidences with local characteristics, such as urban addresses, postal codes, and telephone numbers as found in Web pages. The geospatial evidences are implicitly related to places, so that the contents of a page, or parts of it, can be correlated to an urban geographic location. Thus, search engines can, for instance, use such information to retrieve pages that are related to services and activities in a certain location or close to it.

Therefore, the main contributions of this thesis are (1) the characterization of urban addresses contained in Web pages as sources of geospatial evidences and definition of patterns for their recognition and extraction, (2) the definition of *OnLocus*, an ontology of urban place that helps in the process recognizing and extracting geospatial evidences from Web pages, (3) the creation of a database for recognition of Brazilian places, based on *OnLocus*, (4) the proposal of a strategy for geographic categorization of a Web page, or parts of it, within a country's territorial divisions, and (5) the evaluation of the quantitative and qualitative characteristics of urban addresses that are found in the pages of the Brazilian Web. All of these contributions have been validated through experimentation, using real data from a set of 4 million Web pages. As an additional result, it was possible to obtain a snapshot of the usage of addresses in pages from the Brazilian Web and, consequently, to better understand how to geocode

them. Results of this thesis open a range of perspectives for new types of applications, such as, for instance, the use of navigational links based on geographic location, geographic classification of Web pages, Web-based geospatial data mining, and semantic annotation of pages.

Sumário

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Consultas que incluem pelo menos um termo relacionado a geografia, como nomes de lugar e feições naturais (ex., "praia", "serra"), são hoje um subconjunto significativo das consultas submetidas às máquinas de busca [SaKo04, SMCC06]. Nessas consultas, a especificação do contexto geográfico freqüentemente requer o uso de relações espaciais referentes a distância ou continência, mas esses termos não são entendidos pelas máquinas de busca tradicionais. A inexistência nas máquinas de busca de facilidades para formulação de consultas espaciais, onde expressões espaciais, como "próximo a", "dentro de" ou "ao lado de" e suas variações semânticas, possam ser usadas como uma condição restritiva, faz com que as respostas incluam resultados indesejáveis. Um exemplo desse problema seria uma consulta para se encontrar um hotel em Belo Horizonte, expressa pelos termos "hotel", "em" e "belo horizonte", onde o termo "em" semanticamente significa "dentro de". Essa consulta normalmente é feita digitando-se apenas os termos "hotel" e "belo horizonte", o que nem sempre captura o contexto geográfico desejado. Por exemplo, a execução dessa consulta no Google[1] apresenta como primeira resposta a página do Hotel Belo Horizonte no estado do Rio Grande do Norte. Acrescentando o termo "em", a página do Hotel Belo Horizonte continua sendo retornada. Formulando a mesma consulta como uma frase "hotel em belo horizonte", páginas que continham hotéis de Belo Horizonte, mas que não continham a frase completa, deixam de ser retornadas. Um outro exemplo desse problema ocorre quando o nome do lugar procurado é usado em diversos contextos ou existe em diversos lugares. A busca de um hotel na cidade de São Francisco formulada usando-se os termos "hotel" e "são francisco" e submetida ao Google Brasil, com a opção "só páginas do Brasil", retorna como primeira resposta a página de um hotel em

[1] http://www.google.com

São Francisco, EUA. Da mesma forma, a consulta “hotel” “são vicente” retorna páginas de hotéis em São Vicente na Ilha da Madeira (Portugal) e em São Vicente, São Paulo. Entretanto, existe também a cidade de São Vicente nos estados do Ceará, Rio Grande do Norte, Pernambuco e Minas Gerais, ou seja, o contexto geográfico determinado pelos termos da consulta não é considerado.

Isto ocorre porque, tradicionalmente, a busca na Web envolve o uso de palavras-chave. Quando, em uma máquina de busca tradicional, o nome de um local é pesquisado, as páginas da Web que incluem esse nome em seu texto são recuperadas, ainda que, no contexto da página, ele não signifique um local. As máquinas de busca tradicionais ignoram o escopo geográfico existente nas páginas da Web retornando, normalmente, resultados que não são geograficamente relevantes [BCGG99, DiGS00, JPRS02]. Como conseqüência, muitas vezes só são recuperadas páginas que contenham termos que casem exatamente com termos usados na expressão da consulta. Páginas referentes ao mesmo lugar, mas que usam nomes alternativos, lugares que são próximos ou equivalentes, ou mesmo que estejam fisicamente dentro de uma determinada região, provavelmente não serão encontradas. Além disso, o contexto geográfico existente nas páginas é muitas vezes implícito, podendo ser inferido pela existência de um número de telefone ou CEP.

McCurley [Mccu01] realizou alguns experimentos para verificar qual a fração de páginas da Web que apresentava contextos geográficos reconhecíveis, nesta tese denominados de evidências geo-espaciais, e os resultados mostraram que 4,5% das páginas continham um código postal (*Zip-Code*), 8,5% continham um número de telefone e 9,5% continham pelo menos um dos dois. Segundo o autor, não se pode dizer que o contexto geográfico seja limitado a 10% da Web porque existem outros indicadores para inferir esse contexto implícito em suas páginas que não foram considerados. As evidências geo-espaciais asseguram uma correlação com uma localização geográfica pois, apesar de não possuírem coordenadas geográficas explícitas, podem ser facilmente convertidas em dados posicionais [ArSO00].

O potencial da busca geográfica vem sendo reconhecido tanto no meio acadêmico quanto no meio comercial. Muitas máquinas de busca e provedores, como *Google*, *Yahoo*! e *MSN*, estão empenhados em criar ferramentas que unam a tecnologia de busca à prestação de serviços. São oferecidos serviços de busca que listam estabelecimentos

comerciais, localizando-os em mapas e imagens de satélite. Outros serviços ajudam o usuário, ou um viajante, a localizar endereços em mapas ou navegar sobre um mapa ou imagem de satélite. Entretanto, para oferecer esse tipo de informação, os serviços de interesse dos usuários, normalmente provenientes de bancos de dados comerciais, já estão previamente cadastrados e podem ser localizados em bancos de dados geográficos locais que funcionam como "Páginas Amarelas" na Web.

A busca por informação local (busca local[2]) vai além do uso de páginas amarelas ou mapas na Web porque ela deve refletir todo o leque de atividades que as pessoas participam. A definição de busca local inclui aspectos comerciais e não comerciais como, por exemplo, nas consultas "me mostre os pontos turísticos que estão localizados até 1000 metros daqui" ou "me mostre um restaurante próximo daqui". A Web tem potencial para prover acesso à informação local de forma mais eficiente do que os meios de distribuição tradicionais tanto para o consumidor local quanto para um viajante. Entretanto, as atuais ferramentas de busca local sub-utilizam o potencial que a Web tem a oferecer como um importante repositório de informações geográficas [BLMS03, LBCM05]. Dados sobre cidades vêm constantemente sendo acumulados nas páginas da Web, contendo informações relevantes para a vida cotidiana. Essas páginas são denominadas *páginas de interesse local*, ou seja, páginas relevantes dentro de um determinado limite geográfico ou de interesse para usuários próximos de uma localização, como páginas de restaurantes, teatros, imobiliárias e serviços em geral [BCGG99, DiGS00, GrHL03]. A essa coleção de páginas que satisfaz uma busca por informação de conteúdo local denominamos nesta tese de *Web local*. Assim, para garantir o acesso eficiente a essas páginas, é desejável que uma máquina de busca possa identificar uma página da Web com significado geográfico e possa definir também qual a localização geográfica a ela associada. Por isso, esta tese considera o endereço, dentre as diversas evidências geo-espaciais consideradas na literatura (ver Capítulo 4), como a evidência geo-espacial mais adequada às aplicações de busca local, especialmente considerando a busca de locais intra-urbanos, por representar a localização física de serviços e atividades presentes em páginas da Web.

Para uma máquina de busca encontrar páginas compatíveis com o escopo geográfico de uma consulta, é preciso que as páginas tenham passado por um pré-processamento,

[2] Busca por conteúdo local, direcionada por critérios geográficos como, por exemplo, "restaurante a 500m de um determinado local", "hotéis em Belo Horizonte".

onde as evidências geo-espaciais de características locais sejam localizadas no texto das páginas e depois convertidas em coordenadas geográficas (geocodificação). Se as coordenadas geográficas atribuídas a uma página estiverem contidas na área definida pelo escopo geográfico da consulta, a página é considerada compatível. Assim, no processamento de uma consulta, a localização pode ser considerada e a ordenação dos resultados poderia também levar em consideração a relevância geográfica, o que daria uma precisão maior aos resultados da busca para uma classe de consultas [BCGG99, DiGS00, JPRS02]. No entanto, esse pré-processamento não é trivial, porque as evidências geo-espaciais não só aparecem em qualquer posição dentro do texto não estruturado das páginas, mas também ocorrem, em diversos formatos. Por outro lado, ignorar a presença de endereços em páginas da Web e a possibilidade de sua geocodificação faz com que uma busca por conteúdo local fique dependente e limitada aos serviços e atividades cadastrados e utilizados pelas ferramentas de busca local. Comparado com outras evidências geo-espaciais, a presença de endereços em páginas da Web é ainda pouco explorada, existindo poucos estudos a respeito de suas características e incidência.

O trabalho descrito nesta tese considera a Web como uma rica fonte de conteúdo local e o endereço como uma evidência geo-espacial fundamental no contexto da *Web local*. Assim, esta tese foi direcionada pelas seguintes hipóteses:

(a) Inferências espaciais na Web

- Informações sobre locais podem ser recuperadas através de textos em páginas da Web e são fortemente centradas na noção de endereços.

- Endereços postais podem ser reconhecidos em páginas da Web, pois geralmente estão presentes, tanto de maneira completa ou incompleta, em páginas que oferecem algum tipo de serviço, e são uma evidência altamente confiável para reconhecimento de lugares.

- Telefones normalmente estão presentes em páginas de serviços, atividades e eventos para possibilitar o contato de clientes. Números de telefone acompanhados do código de área também são uma evidência confiável para reconhecimento de lugares.

- Códigos de Endereçamento Postal, quando presentes em páginas da Web, são um indicador fortíssimo da presença de endereços e, conseqüentemente, uma evidência confiável para reconhecimento de lugares.
- Consultas geográficas, submetidas a máquinas de busca, são relativamente freqüentes [SaKo04, SMCC06], justificando métodos para otimizá-las.

(b) Cognição espacial na Web

- As pessoas se comunicam e interagem através de uma linguagem qualitativa, dando direções sem detalhes precisos. Conseqüentemente, o uso de expressões que denotam relações espaciais, como proximidade e continência, são freqüentes na descrição de localizações incluídas no texto das páginas encontradas na Web [MGGF03, Zhan03a].
- As pessoas usam pontos de referência para indicar, de forma aproximada, a localização de locais de interesse. Referências a lugares, citadas em páginas da Web, tendem a destacar a proximidade com lugares conhecidos localmente.
- A localização de um lugar de interesse pode ser determinada, de maneira aproximada, a partir da localização de um ponto de referência e de uma interpretação dada às expressões em linguagem natural que denotam relacionamentos espaciais.

1.2 Objetivo e Contribuições

O surgimento de serviços como *Google Earth*[3] e *Virtual Earth*[4] que mostram o planeta inteiro em imagens de satélite, com possibilidade de aproximação até o nível local, abre inúmeras possibilidades de integração entre a busca local e sua visualização. No entanto, as facilidades para busca local oferecidas ainda são insuficientes. Os serviços de busca local disponibilizados por provedores ou pelas máquinas de busca atuais funcionam de forma limitada para regiões do mundo fora da América do Norte e Europa e, mesmo lá, as informações mais precisas só estão disponíveis para as grandes cidades. Além disso, qualquer tipo de busca local requer a habilidade de se efetuar busca por proximidade, que é a capacidade de encontrar informação associada com

[3] http://earth.google.com
[4] http://local.live.com

alguma localização que esteja a uma certa distância de um local de interesse. Para esse tipo de busca, a Web é um vasto repositório de informações locais e informações geográficas. Esse conteúdo está presente na Web independentemente da deficiência dos mecanismos para acessá-lo e transformá-lo em informação disponível.

Dentro desse contexto, esta tese propõe uma abordagem para reconhecer e extrair evidências geo-espaciais de características locais, como endereços, códigos postais e telefones, presentes nas páginas da Web que, por exemplo, poderiam ser utilizadas pelas máquinas de busca para recuperação de páginas referentes a serviços e atividades em uma determinada localidade ou próximos a ela. Além disso, como veremos no Capítulo 6, inúmeras aplicações adicionais poderiam se beneficiar dessa abordagem. Entre elas podemos citar: classificação geográfica de páginas da Web, uso de *links* de navegação baseados em localização geográfica, mineração de dados geo-espaciais em páginas da Web e anotação semântica das páginas.

Nesta tese, as evidências geo-espaciais representam localizações implícitas capazes de correlacionar o conteúdo de uma página, ou de parte dela, a uma localização geográfica urbana. Dessa forma, a tese tem como foco a *Web local*, propondo estratégias, apoiadas em uma ontologia de lugar urbano, que permitam reconhecer, extrair e geocodificar evidências geo-espaciais sob a forma de endereços encontradas em páginas da Web.

Assim, as principais contribuições desta tese são:

- Definição de uma ontologia de lugar urbano, denominada *OnLocus*, para auxílio no processo de reconhecimento e extração de evidências geo-espaciais de páginas da Web que permitam a geocodificação automática dessas páginas;
- Criação de uma base de conhecimento de lugares (*gazetteer*[5]) do Brasil baseada na ontologia de lugar urbano e que foi utilizada para implementação de um sistema de localização espacial denominado *Locus* [SDBD05, SDBD04, Souz05];
- Caracterização dos endereços presentes em páginas da Web como fontes de evidência geo-espacial [BLMS03, LBCM05] e definição de padrões para o seu reconhecimento e extração;

[5] *Gazzetter* é um dicionário geo-espacial de nomes geográficos (Capítulo 2)

- Proposta de uma estratégia para categorização geográfica de uma página, ou de partes dela, dentro da divisão territorial (País/Estado/Cidade);
- Avaliação das características qualitativas e quantitativas dos endereços presentes nas páginas da Web brasileira.

A seguir, na Seção 1.3, é apresentada uma visão geral do processo de reconhecimento, extração e geocodificação de evidências geo-espaciais sob a forma de endereços encontrados em páginas da Web, onde pode ser constatado o papel essencial da ontologia *OnLocus* e de um *gazetteer* no reconhecimento e na validação dos endereços extraídos. Esse processo foi executado na Web brasileira para validação da proposta desta tese. Como será visto, o conhecimento adquirido sobre as características dos endereços na Web brasileira serviram de subsídio para a definição de padrões de reconhecimento e extração de endereços.

A partir de agora, toda a referência a lugares e ontologia de lugar, no texto, será específica a lugares urbanos e conceitos associados, evitando-se a repetição da qualificação "urbano" para não sobrecarregar o texto.

1.3 Visão Geral

Como anteriormente mencionado, esta tese considera o endereço um elemento fundamental para a busca local, reconhecendo a existência de endereços completos, incompletos e parciais, conforme encontrados em páginas da Web. Um *endereço completo* é aquele que possui uma parte básica (tipo de logradouro, nome de logradouro e número de imóvel) seguida de todos os componentes identificadores de localização - telefone, CEP, cidade e estado. Um *endereço incompleto* possui a parte básica seguida de pelo menos um dos identificadores de localização e um *endereço parcial* só possui identificadores de localização - CEP ou telefone (ver Capítulo 4).

A Figura 1.1 apresenta as principais fases envolvidas no processo proposto para o reconhecimento, extração e geocodificação de endereços existentes em páginas da Web. Uma descrição geral de cada uma das fases é apresentada a seguir.

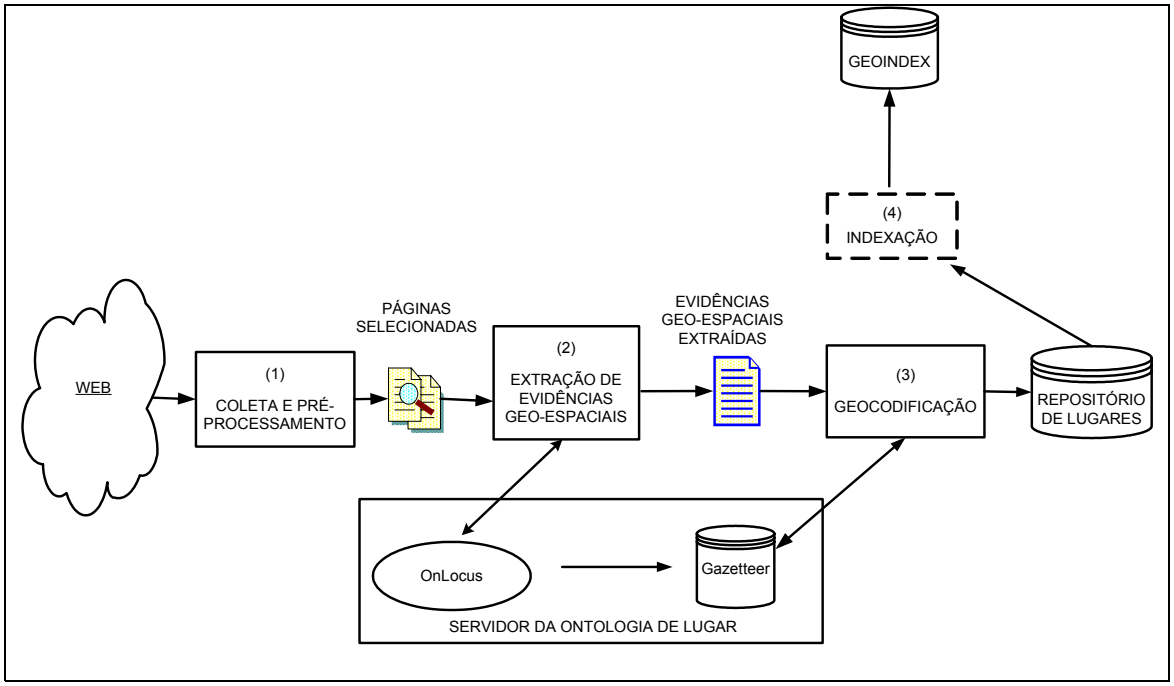


Figura 1.1 – Visão geral do processo de reconhecimento, extração e geocodificação de evidências geo-espaciais sob a forma de endereços completos, incompletos e parciais

Inicialmente, as páginas da Web são coletadas e pré-processadas (1). Esse pré-processamento tem por objetivo selecionar documentos no formato HTML (denominados de páginas) e normalizar o conjunto de *tokens* e delimitadores, reduzindo a complexidade dos padrões de reconhecimento. A seleção de documentos no formato HTML se deve ao fato deste ser o formato de maior presença na Web. No pré-processamento são removidos documentos repetidos (identificados com base na sua URL) e os que não estejam no formato HTML como, por exemplo, documentos nos formatos PDF (*Portable Document Format*), PS (*PostScript*), DOC (*MS-Word*)) ou PPT (*MS-Powerpoint*). Os documentos HTML são pré-processados para substituição dos seus marcadores (*tags*) por um marcador especial, remoção dos acentos ortográficos, substituição de caracteres de controle, tais como TAB, FF e CR, por espaço e eliminação de espaços consecutivos.

Após a coleta e pré-processamento, as páginas passam para a fase seguinte, onde através de padrões e regras definidas com base na ontologia de lugar *OnLocus*, evidências geo-espaciais em potencial como, endereço, CEP e número de telefone, são reconhecidas, extraídas e armazenadas em um repositório (2). O processo de reconhecimento dos componentes que formam um endereço (completo, incompleto ou parcial) segundo os padrões definidos (ver Capítulo 4) é chamado de *geoparsing*. Nele

só são reconhecidos endereços com presença de pelo menos um dos identificadores de localização (CEP, número de telefone com código de área, nome de cidade ou estado), o que possibilita a identificação da cidade à qual o endereço está associado. É importante destacar que no processo de reconhecimento de um endereço, o idioma e a cultura local devem ser considerados. Apesar de universalmente os endereços serem formados por praticamente os mesmos componentes, a seqüência com que eles aparecem variam de um país para outro. Países que usam línguas romanas, como português, espanhol e italiano, normalmente usam o tipo do logradouro e o nome do logradouro antes do número do imóvel. Já nos países de língua germânica, o nome e o tipo do logradouro são concatenados formando uma única palavra. Nos Estados Unidos é comum o uso do número do imóvel e do prefixo de direção (N, NW, S, etc.) antes do nome do logradouro. A Figura 1.2 ilustra algumas dessas características encontradas em endereços descritos em vários idiomas e para diferentes países.

Itália	Viale Manlio Gelsomini 28, 00153 Rome, Italy
Áustria	Siebenbrunnengasse 44, A-1050 Vienna, Austria
EUA	702 H Street NW, Suite 300, Washington DC 20001, USA.
Brasil	Rua Alvarenga, 2331, Butantã 05509-006, São Paulo/SP, Brazil
França	22 rue des rasselins 75020 Paris, France

Figura 1.2 – Endereços dos escritórios do Greenpeace em vários países
Fonte: http://www.greenpeace.org

O resultado do *geoparsing* é um endereço estruturado extraído e armazenado em um repositório temporário. O endereço estruturado passa então para a etapa de geocodificação (3).

Antes da fase de geocodificação propriamente dita, é verificado se os termos extraídos, como CEP, número de telefone com DDD ou nome de cidade e estado possuem uma correspondência no *Locus* ou seja, se esses termos extraídos quando submetidos ao *Locus*, encontram um CEP, um "DDD + prefixo" ou uma cidade e estado já cadastrados. Existindo tal correspondência, a localização é reconhecida e validada, possibilitando, assim, a geocodificação do endereço. A fase de geocodificação (*geocoding*) compreende a localização de pontos na superfície da terra a partir de informações alfanuméricas [DaFB03], envolvendo duas etapas: casamento (*matching*) e localização (*locating*). Na etapa de casamento é estabelecida uma correspondência entre o endereço identificado e uma entidade geográfica (por exemplo, uma rua, cidade ou

estado) existente no *gazetteer* do *Locus* (Capítulo 3). Na etapa de localização é feita a atribuição de coordenadas geográficas (latitude/longitude) a esse endereço (Figura 1.3).

A estratégia de geocodificação de endereços adotada nesta tese considera uma infra-estrutura de endereçamento composta por entidades geográficas pontuais, representando endereços individuais, e entidades geográficas lineares, representando trechos de logradouro com faixa de numeração, e utiliza os relacionamentos de localização definidos na ontologia *OnLocus* - *localização exata*, *localização aproximada* e *localização genérica* (ver Capítulo 3). A *localização exata* acontece quando o endereço extraído encontra um endereço correspondente no *Locus*. Assim, as coordenadas do endereço encontrado são atribuídas ao endereço extraído. A *localização aproximada* acontece quando o endereço extraído não encontra uma correspondência exata no *Locus*, mas pode ser associado ao número de um imóvel mais próximo, ou quando uma determinada cidade não possui endereços localizados individualmente, mas por faixa de numeração em cada trecho de um logradouro. A *localização genérica* é usada quando nenhuma das situações anteriores for capaz de localizar um endereço. Por conseguinte, é utilizada qualquer informação adicional, geograficamente localizável, como limites do bairro, limite da cidade ou áreas de CEP, que possibilite localizar o endereço da forma mais precisa possível. A localização genérica poderá ser feita de três maneiras: (1) no centróide do polígono que representa a região escolhida, (2) aleatoriamente dentro da região escolhida ou (3) usando o retângulo mínimo envolvente (RME) da região escolhida.

A Figura 1.3 resume o processo de geocodificação mostrando a seqüência das etapas desde o *geoparsing* do endereço até a sua localização.

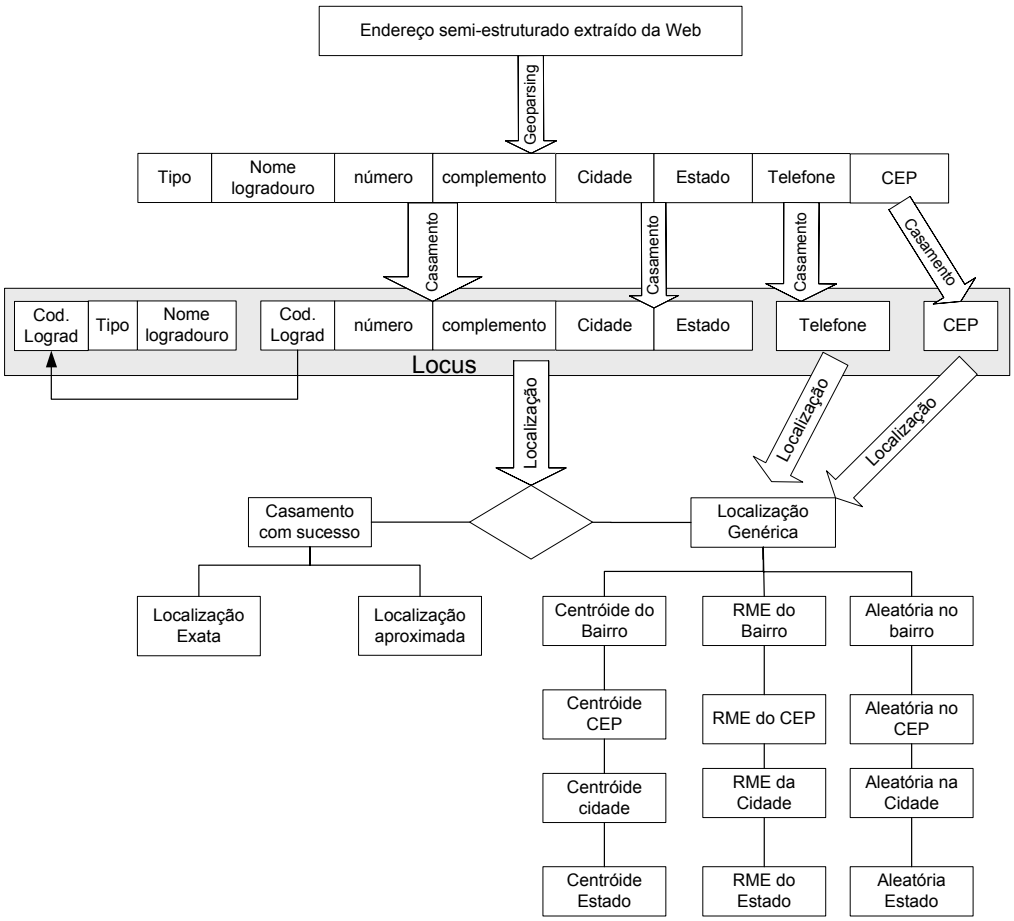


Figura 1.3 – Visão geral do processo de geocodificação de endereços

Como resultado da geocodificação, temos um *Repositório de Lugares* contendo, para cada endereço extraído e validado, a URL da página, o padrão usado na extração, os termos extraídos, a posição inicial e final dos termos na página, o nome da cidade, o nome do estado, as coordenadas geográficas do RME da cidade e, quando for possível, as coordenadas geográficas do endereço juntamente com o tipo de localização usado na geocodificação. Dessa forma, o Repositório de Lugares proveria a informação necessária para a etapa de indexação (4), onde um índice espacial (GeoIndex) seria gerado para permitir o acesso a páginas que possuam algum contexto geográfico. Outros índices, que permitissem outros tipos de busca espacial, também poderiam ser criados com os dados extraídos, como, por exemplo, um índice para agrupar as páginas pela subdivisão territorial do País, que utilizasse a informação de cidade e estado obtidas. Este índice traria vantagens em relação à indexação tradicional das máquinas de busca, porque como são usados vários tipos de evidência geo-espacial para a atribuição do nome de uma cidade, já que nem sempre os nomes de cidade e estado estão presentes no texto das páginas da Web.

A seguir, é apresentado um exemplo para ilustrar o processo de reconhecimento, extração e geocodificação de endereços, tendo como base o fragmento de uma página da Web brasileira apresentado na Figura 1.4.

Figura 1.4 – Presença de endereços em páginas da Web brasileira com a identificação de seus componentes.

Para o reconhecimento de um endereço foram definidos dezoito padrões (ver Capítulo 4) que, conforme observado em nossos experimentos, refletem as diversas maneiras de se descrever um endereço nas páginas da Web brasileira. Os padrões definidos foram traduzidos em expressões regulares e implementados utilizando a linguagem PERL [WaCS96]. Dentre esses dezoito padrões, foram selecionados seis que

demonstraram, através de experimentos, serem suficientes e eficazes no reconhecimento de um endereço (ver Capítulos 4 e 5). A Figura 1.4 mostra um exemplo de endereço reconhecido em uma página da Web brasileira. Neste exemplo, são utilizados quatro dos padrões definidos e também o padrão que reconhece todos os componentes encontrados no endereço considerado. Considerando que o tamanho do complemento foi experimentalmente limitado a 30 posições, a Figura 1.5 exemplifica a estratégia adotada para o reconhecimento do endereço indicado na Figura 1.4.

O reconhecimento dos componentes (*geoparsing*) do endereço "Rua Gamboa de Cima, 236 - Gamboa - Cep 40080-060 Fone: (71) 337-1303" pode ser feito utilizando-se os seguintes padrões: (1) *Endereço Básico* + *CEP*, (2) *Endereço Básico* + *Telefone*, (3) *CEP*, (4) *Telefone* e (5) *Endereço Básico* + *CEP* + *Telefone*. Por exemplo, se o padrão utilizado for *Endereço Básico* + *CEP* será reconhecido o endereço "Rua Gamboa de Cima, 236", o complemento "- Gamboa -" e o CEP "40080-060". Já se o padrão utilizado for *Endereço Básico* + *Telefone* será reconhecido o endereço "Rua Gamboa de Cima, 236", o complemento "- Gamboa - Cep 40080-060" e o telefone "(71) 337-1303". Neste caso, uma vez que "- Gamboa - Cep 40080-060" ocupa menos de 30 posições e o padrão utilizado procura por um telefone após o endereço, o CEP será considerado como um complemento. O endereço completo é reconhecido pelo padrão *Endereço Básico* + *CEP* + *Telefone*.

Após o *geoparsing* do endereço, cada termo reconhecido, ou um conjunto deles, que representa um componente do endereço é armazenado em um repositório temporário para posterior validação. Considerando que o padrão utilizado foi *Endereço Básico* + *CEP* + *Telefone*, são armazenados no repositório: "Rua" como um tipo de logradouro, "Gamboa de Cima" como nome do logradouro, "236" como o número do imóvel, "- Gamboa -" como complemento, "40080-060" como o CEP do endereço, "71" como o código de área do telefone, "337" como o prefixo do número de telefone e "1303" como o sufixo, além da URL da página e das posições inicial e final onde os termos foram encontrados.

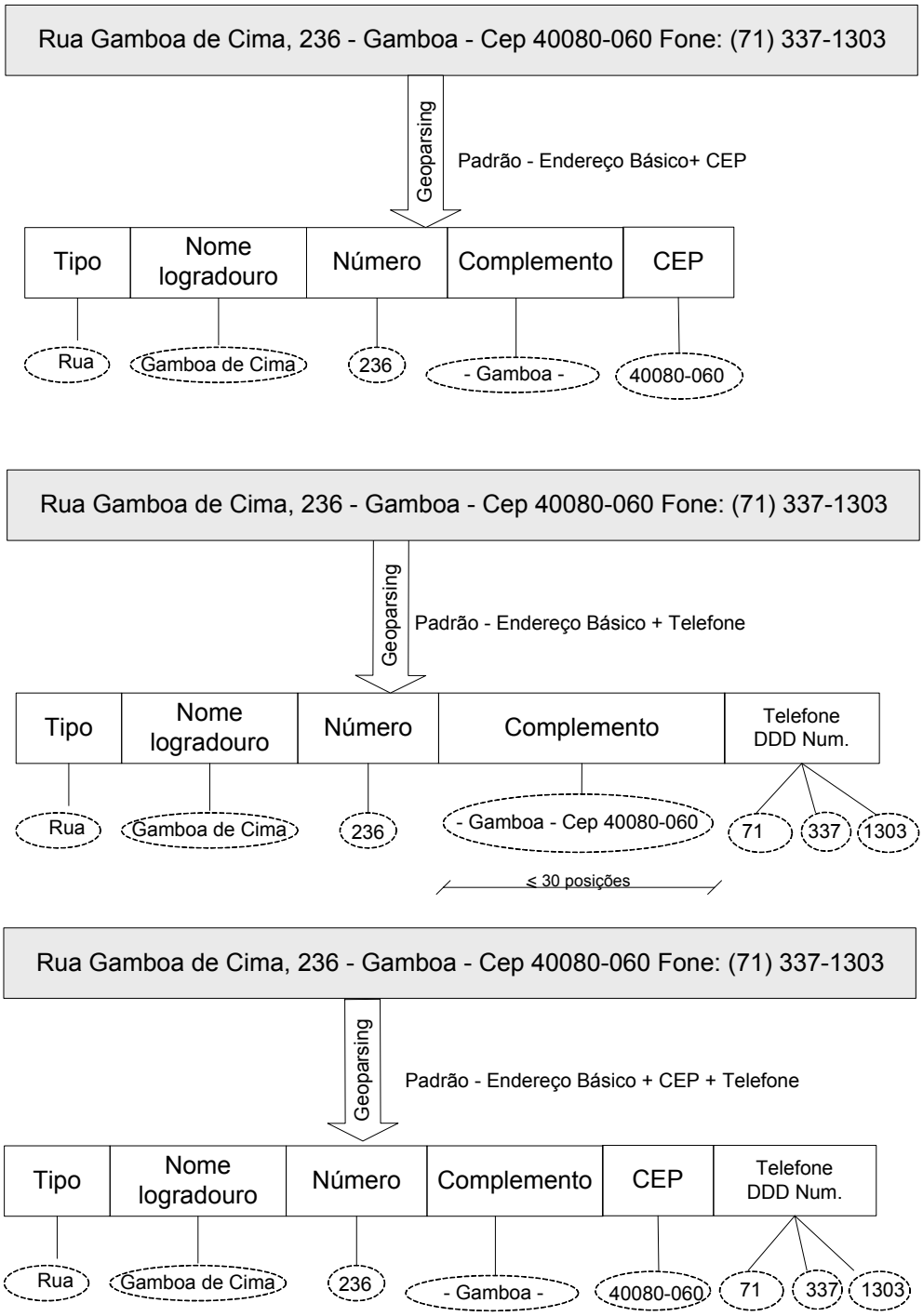


Figura 1.5 – Exemplos de *geoparsing* usando diferentes padrões.

Esse endereço armazenado é considerado um endereço em potencial até que tenha pelo menos um de seus identificadores de localização validado pelo *Locus*. Para o identificador de localização *Telefone*, é verificado se o "DDD + prefixo" existe no *gazetteer* do *Locus* e, quando existente, são recuperados o nome da cidade e do estado correspondentes. Para o identificador *CEP*, o mesmo procedimento é executado.

Quando em um mesmo endereço existem dois identificadores, como no caso do exemplo da Figura 1.4, serão considerados apenas o nome da cidade e do estado obtidos através do *CEP*. Os resultados experimentais apresentados no Capítulo 5 mostram que o *CEP* é o identificador de localização mais confiável. No nosso exemplo, o CEP 40080-060 foi encontrado no *gazetteer* do *Locus* correspondendo à cidade de Salvador, no estado da Bahia. O registro no repositório temporário é então acrescido dessa informação juntamente com as coordenadas geográficas do RME de Salvador.

Validada a cidade à qual o endereço está associado, passa-se à fase de geocodificação do endereço. Com o endereço estruturado e validado, é feita nova consulta ao *gazetteer* do *Locus* para verificar a sua existência. Se no entanto, não existir qualquer informação que possibilite a sua geocodificação, existindo apenas o limite geográfico da cidade, a localização do endereço será *genérica* podendo ser atribuídas as coordenadas geográficas do centróide do município, do RME ou de um ponto aleatório dentro do polígono do município. Por outro lado, se existir um endereço correspondente no *gazetteer* do *Locus*, a localização será exata e as coordenadas geográficas encontradas são então atribuídas ao endereço extraído. A Figura 1.6 ilustra a visualização do Flat Solar da Gamboa, indicado na Figura 1.4, de acordo com as coordenadas do seu endereço encontradas no *gazetteer* do *Locus*. É importante ressaltar que o *gazetteer* do *Locus* [SDBD05, Souz05,] constitui uma implementação da ontologia *Onlocus* que utiliza dados dos Correios, da ANATEL e do IBGE e é utilizado como uma base de conhecimento de locais brasileiros.

Após a geocodificação, o Repositório de Lugares é atualizado com os dados extraídos, as coordenadas geográficas do endereço e com o tipo de localização que foi utilizado na geocodificação, no caso do exemplo, o tipo poderia ser *genérico* ou *exato*. Para busca local, onde a localização de um endereço é fundamental, apenas as localizações *exata* e *aproximada* são consideradas. Entretanto, para categorização das páginas dentro da divisão territorial de um país, qualquer um dos três tipos de localização pode ser considerado.

Figura 1.6 – Visualização do Flat Solar da Gamboa de acordo com as coordenadas geográficas do seu endereço fornecidas pelo *Locus*.

1.4 Organização da Tese

O restante desta tese está organizado da seguinte forma. O Capítulo 2 descreve os principais conceitos utilizados para sua fundamentação (ontologias, *gazetteer* e relacionamento espacial) e apresenta os trabalhos relacionados. O Capítulo 3 descreve a ontologia de lugar – *OnLocus* proposta para auxiliar no reconhecimento de lugares em páginas da Web. O Capítulo 4 discute as evidências geo-espaciais encontradas nas páginas da Web e as formas de extraí-las. No Capítulo 5 são apresentados os resultados experimentais obtidos em páginas da Web brasileira e, por fim, o Capítulo 6 apresenta as conclusões da tese e sugere temas para trabalhos futuros.

Capítulo 2

"O espaço é o
mais interdisciplinar dos objetos concretos"
Milton Santos

Conceitos e Trabalhos Relacionados

Este capítulo descreve alguns conceitos básicos utilizados na tese, como ontologias, *gazetteers* e relacionamentos espaciais, e apresenta os trabalhos relacionados.

2.1 Ontologias

Ontologia é a ciência que estuda o ser e suas propriedades. O termo ontologia vem do grego – *ontos* - que significa ser. O significado de ontologia usado na área de computação veio dos filósofos gregos, especialmente Aristóteles, que investigaram as propriedades do mundo, dos objetos e como nós as percebemos. Em computação, pesquisas em ontologia tiveram sua origem nas comunidades de Inteligência Artificial (IA) e de Representação do Conhecimento [SCPV04]. Neste contexto, uma ontologia é um formalismo que permite especificar um vocabulário relativo a um determinado domínio. Esse vocabulário define entidades, classes, propriedades, predicados, funções e as relações entre estes componentes no contexto de algum domínio específico [Grub92, Roch03].

2.1.1 Ontologias em Computação

Embora o termo ontologia seja hoje em dia freqüentemente usado e sua importância reconhecida, não existe um consenso sobre seu exato significado [Grub95]. A comunidade de IA vê ontologias como teorias de lógica formal através das quais não só são definidos termos e relacionamentos, mas também o contexto no qual os termos estão aplicados, e fatos e relacionamentos inferidos. Por outro lado, outras comunidades, como a de bancos de dados, vêem ontologias mais como modelos de objetos, taxonomias e esquemas, não expressando explicitamente restrições importantes.

Ontologias lingüísticas como *WordNet*[6] ou tesauros expressam vários relacionamentos entre conceitos (ex., sinônimos, antônimos, "é-um", "parte-de") mas não descrevem explícita e formalmente o que os conceitos significam [Khan00].

Spaccapietra et al. [SCPV04] classificam ontologias em dois tipos: a taxonômica e a descritiva. A ontologia taxonômica abrange a primeira geração de ontologias, ainda muito usada, e está focada na definição de termos e na sua organização em hierarquias de generalização/especialização [SmSm77,TsLo82], enriquecidas com relações semânticas comumente usadas em lingüística, como sinônimo e antônimo. Uma ontologia taxonômica define um vocabulário de referência. Esse tipo de ontologia é relativamente fácil de ser usado. *WordNet* é um exemplo de ontologia taxonômica.

As ontologias descritivas e os esquemas conceituais de bancos de dados compartilham o mesmo esforço de modelar algum domínio ou atividade. No entanto, enquanto as ontologias têm sido tradicionalmente consideradas um meio de descrever o mundo, os esquemas de bancos de dados também auxiliam o gerenciamento dos dados que representam o mundo de interesse [SCPV04].

Como uma ontologia provê um conjunto de conceitos e termos para descrever algum domínio, ela representa a estrutura básica sobre a qual uma base de conhecimento poderá ser construída. A base de conhecimento faz uso desses termos para representar o que é verdadeiro sobre algum mundo real ou hipotético [SwTa99]. Ontologias com instâncias são freqüentemente referenciadas como bases de conhecimento [SCPV04].

De acordo com o tipo de conhecimento que uma ontologia descreve, ela pode ser classificada como genérica ou dependente do domínio. Uma ontologia genérica especifica conceitos gerais definidos independentemente do domínio da aplicação, podendo ser usada em diferentes domínios. OpenCyc[7], SUMO[8] (Suggest Upper Merged Ontology) e *WordNet* são exemplos de ontologias genéricas. Uma ontologia dependente do domínio é dedicada a um domínio particular, capturando o conhecimento de especialistas referente a uma determinada aplicação, como, por exemplo, uma ontologia biomédica [Khan00, Roch03].

[6] http://wordnet.princeton.edu
[7] http://www.opencyc.org
[8] http://www.ontologyportal.org

2.1.2 Linguagens para Especificação de Ontologias

Uma ontologia precisa ser especificada em alguma linguagem formal e implementada para que seja entendida, compartilhada e exportada entre diferentes usuários. Existem várias linguagens de especificação de ontologias, com graus de formalismo diferentes. Algumas são baseadas em XML (*eXtensible Markup Language*) como XOL (*XML-Based Ontology Exchange Language*) [KaCT99], SHOE (*Simple HTML Ontology Extensions*) [HeHL99] e RDF (*Resource Description Framework*) [LaSw99], outras em lógica de descrição como KIF (*Knowledge Interchange Format*) [GeFi92], existindo ainda aquelas que combinam XML e lógica de descrição, como OIL [FHHM01] e DAML+OIL [HoHa00].

RDF é um arcabouço desenvolvido pelo W3C (*World Wide Web Consortium*) [W3CS01] para a representação de informação na Web utilizando a sintaxe XML. RDF Schema (RDFS) especifica os vocabulários, em RDF, usados para a representação de informação na Web por meio de classes, propriedades, tipos, intervalos e domínios. RDF/RDFS é um padrão voltado para a Web semântica recomendado pelo W3C para representação de informação [BeHL01].

DAML+OIL é uma combinação de DAML (DARPA *Agent Markup Language*) e OIL (*Ontology Inference Layer*) que estende a linguagem RDF/RDFS adicionando mais expressividade. Ela tem uma sintaxe RDF/XML baseada no paradigma de *frames* e descreve a estrutura de um domínio (esquema) em um estilo orientado a objetos [Zhan03].

A *Web Ontology Language* (OWL) [McHa04] é uma linguagem de marcação para a descrição de ontologias. Ela foi projetada pelo *Web Ontology Working Group*[9]. Como RDF, é um padrão recomendado pelo W3C para a Web Semântica [BeHL01], tanto para especificação como para intercâmbio de ontologias. As principais influências na especificação da linguagem OWL vieram da lógica de descrição, das linguagens RDF/RDFS e DAML+OIL, e do paradigma de *frames* [HoPH03]. Nas linguagens baseadas em *frames*, cada conceito é descrito como uma classe e cada classe é descrita em um *frame*. Em mais detalhes, classes são definidas por atributos (*slots*) e

[9] `http://www.w3.org/2001/sw/WebOnt/`

estruturadas de acordo com o relacionamento com as subclasses dentro de um grafo ou de uma taxonomia [Roch03].

Estudos comparativos de várias linguagens usadas para descrever ontologias são apresentados em [CoFG03], [HoPH03], [Roch03] e [SCPV04]. Há várias propostas de ferramentas para o desenvolvimento de ontologias como, por exemplo, Protégé 2000 [NSDC01], OilEd [BHGS01], KAON [BEHH02], OntoEdit [SEAS02], Ontolingua [Grub92a], Ontosauros [SPKR96], OntoBroker [DEFS99], WebODE [ACFG01], WebOnto [DoMC99] e RACER [HaMo03].

2.1.3 Ontologia de Extração

Embley et al. [ECJL99, Embl04] definem uma ontologia de extração como uma instância de um modelo conceitual que descreve em termos de conjuntos de objetos e relacionamentos, uma determinada aplicação em um domínio de interesse específico. Para cada conjunto de objetos representado nessa ontologia existe um *data frame* que define seu conteúdo potencial em termos de expressões regulares e palavras-chave. Assim, uma ontologia de extração é uma instância de um modelo conceitual associado a regras de extração, que funciona como um *wrapper* (extrator de dados) para um domínio de interesse. Quando uma ontologia de extração é aplicada a páginas da Web, ela identifica objetos e relacionamentos e os associa ao conjunto de objetos e de relacionamentos representado na instância do modelo conceitual daquele domínio. Com essa abordagem, também é identificado o significado semântico de cada termo reconhecido e extraído [Embl04].

2.1.4 Ontologia Geográfica

A maioria das ontologias existentes não considera as características espaciais e temporais da informação [SCPV04]. Uma ontologia relacionada ao espaço geográfico difere de outras porque a topologia e os relacionamentos "todo-parte" assumem um papel de destaque no domínio geográfico. O espaço geográfico pode ser particionado de diversas formas e segundo diferentes escalas e visões [FEAC02]. Objetos geográficos são tipicamente complexos e possuem múltiplas partes. Podem ser conectados ou contíguos, dispersos, fechados ou abertos [SmMa98].

Segundo Spaccapietra et al. [SCPV04], o espaço pode participar de uma ontologia de três formas distintas. Na primeira, o espaço é o próprio domínio, descrito através de

conceitos como os elementos espaciais (ponto, linha e polígono) e os relacionamentos espaciais. Na segunda, o espaço aparece implícito como nas ontologias de domínio geográfico onde as aplicações que lidam com dados geográficos usam o espaço como o local onde as aplicações acontecem (por exemplo, sistemas de rede de transporte). Na terceira, o espaço é usado para enriquecer a descrição de conceitos de uma ontologia, descrevendo por exemplo uma localização espacial. Uma ontologia que permite especificar características espaço-temporais em seus conceitos é chamada de *ontologia espaço-temporal*.

A mereologia [Simo87, Varz03] e a mereotopologia [Smit96, Varz96] estenderam os estudos ontológicos para o domínio espacial. A mereologia descreve as relações entre as partes e o todo e a mereotopologia estendeu a teoria da mereologia com métodos topológicos.

Várias iniciativas têm surgido com o intuito de descrever conceitos geográficos. Entre elas podemos citar SWEET, SUMO e OpenCyc. SWEET[10] (*Semantic Web for Earth and Environmental Terminology*) é uma iniciativa da NASA que provê várias ontologias de alto nível relacionadas às Ciências da Terra como, por exemplo, *Earth Realm*, *Physical Phenomena*, *Space*, *Time*, *Physical Property* e *Biosphere*. As ontologias são descritas na linguagem OWL [McHa04] e têm por objetivo auxiliar a encontrar e usar dados e informações relacionadas às Ciências da Terra presentes em páginas da Web. Os criadores do SWEET acreditam que o entendimento semântico do texto de forma automática é possível através da combinação de ontologias e ferramentas capazes de interpretá-las. SUMO [NiPe01] é uma ontologia que descreve conceitos gerais e faz parte do *IEEE Standard Upper Ontology Working Group*[11]. SUMO possui também ontologias de domínio geográfico, como *Countries and Region*s e *Geography*, fornecendo relacionamentos hierárquicos entre áreas geográficas e suas subdivisões. OpenCyc é uma ontologia de alto nível descrevendo aproximadamente 47.000 conceitos. Um deles é o conceito *Geography*[12] que descreve conceitos relacionados a regiões do mundo, entidades geopolíticas e localização.

[10] http://sweet.jpl.nasa.gov/index.html
[11] http://suo.ieee.org/
[12] http://www.cyc.com/cycdoc/vocab/vocab-toc.html

Esta tese, como se verá no Capítulo 3, usa uma ontologia geográfica de lugar específica para o reconhecimento e extração de evidências geo-espaciais presentes em páginas da Web referentes a atividades ou serviços oferecidos nas cidades. Trata-se de uma ontologia de extração e dependente de domínio e sua definição usa a ferramenta Protégé 2000 que a exporta nas linguagens RDF(S) e OWL.

2.2 *Gazetteers* Digitais

Na tarefa de associar um nome de lugar às páginas da Web, muitos trabalhos propostos na literatura fazem uso de *gazetteers*. Um *gazetteer* digital (ou simplesmente *gazetteer*) é um dicionário geo-espacial de nomes geográficos, ou seja, uma coleção de nomes de lugar associados à sua localização e a outras informações descritivas. Uma localização é definida através de listas de coordenadas geográficas representando linhas, pontos ou áreas. A função básica de um *gazetteer* é informar a localização de um lugar dado seu nome, uma vez que essa é a forma mais natural de se referir a algum lugar. O nome pelo qual um lugar é conhecido é considerado uma referência espacial indireta [Hill00]. Um *gazetteer* é, assim, uma ferramenta para geocodificar essa referência espacial indireta, isto é, associar uma ou mais coordenadas geográficas ao nome de um lugar.

Os objetos em um *gazetteer* possuem três componentes essenciais: o nome do lugar, sua localização (*footprint*) e o tipo que identifica a categoria do lugar. O nome pode admitir valores alternativos, como apelidos, abreviaturas e nomes não oficiais, antigos ou em outras línguas. A localização exige pelo menos uma coordenada geográfica, podendo ser também um polígono ou um retângulo mínimo envolvente (RME). O tipo é um atributo fundamental no reconhecimento de um lugar, ajudando a solucionar ambigüidades nos casos em que um mesmo nome aparece em várias categorias, como, por exemplo, a cidade do Rio de Janeiro e o estado do Rio de Janeiro. Com esses três atributos básicos, um *gazetteer* é capaz de processar ao menos dois tipos fundamentais de consulta: “onde fica esse lugar?” e “o que há nesse lugar?” [Hill00].

Recentemente, *gazetteers* passaram a ser usados para auxiliar a resolução de ambigüidades em páginas da Web [AHSS04, RaBB03, ZWSL05]. Exemplos de *gazetteers* disponíveis na Web são ADL – Alexandria Digital Library [ADLG04], GNIS - Geographic Names Information System [GNIS06], TGN - Getty Thesaurus of

Geographic Names [GTGN00], U.S. Census 2000 [USCB00] e GNS - GEOnet Names Server [GeNS02] .

No entanto, os *gazetteers* disponíveis possuem algumas limitações que impedem sua completa utilização como ferramenta para recuperação de informação geográfica [FuAJ03]. Grande parte dos *gazetteers* trata apenas de relacionamentos hierárquicos e de continência ("parte-de"), ignorando outros tipos de relacionamento espacial como adjacência e sobreposição. Geralmente, os nomes de lugar existentes em um *gazetteer* são associados a objetos bem definidos, sendo incapazes de tratar localizações imprecisas, como "centro de Belo Horizonte". Além das limitações descritas, existe nos *gazetteers* a falta de nomes intra-urbanos, como pontos de referência, nomes de ruas e de monumentos, como também a ausência de outras informações importantes, como número de telefone (códigos de área e prefixos) e códigos postais.

2.3 Raciocínio Espacial

Raciocínio espacial (*spatial reasoning*) é o termo usado para denotar inferências sobre relacionamentos espaciais entre objetos no espaço, usando um subconjunto conhecido de relações espaciais. Este tipo de raciocínio permite fazer predições e diagnósticos. O raciocínio espacial pode ser classificado como quantitativo e qualitativo, dependendo do tipo de informação usada no processo de raciocínio [Rodr02].

O raciocínio espacial quantitativo distingue diversos relacionamentos espaciais, por exemplo, relacionamentos topológicos ou métricos, e é tipicamente formalizado usando um sistema de coordenadas cartesianas e álgebra vetorial. Este processamento quantitativo da informação é claramente distinto da forma como as pessoas interpretam relações espaciais. O raciocínio espacial qualitativo é amplamente utilizado por seres humanos para entender, analisar e tirar conclusões sobre um ambiente espacial. Segundo Frank [Fran96], no pensamento humano sobre o espaço e situações espaciais, prevalece o raciocínio espacial qualitativo. As pessoas pensam e se comunicam a respeito do mundo em termos de conceitos vagos, que são imprecisos ou probabilísticos, como, por exemplo, "centro da cidade", "perto de", "nos arredores de" [MGGF03, Zhan03a]. As pessoas raramente dizem "o restaurante está a 35,93 metros a oeste", por outro lado

fornecem algumas instruções qualitativas como “o restaurante está à direita, a duas quadras da rodovia”.

O raciocínio espacial qualitativo tem sido proposto como um mecanismo complementar de inferência de relacionamentos espaciais desconhecidos [EgSh98, Fran96, MaEg95, MGGF03], sendo baseado na manipulação de um conjunto restrito de símbolos (isto é, relações espaciais) como Norte, Sul, próximo, etc., para os quais tabelas de composição (conjunto de regras lógicas sobre a combinação desses símbolos) facilitam o raciocínio, permitindo a inferência de novos relacionamentos espaciais. Esta tese analisa o conteúdo de páginas da Web a partir de aspectos qualitativos de raciocínio espacial, usando dentre outros vários tipos de relacionamento espacial.

2.4 Relacionamentos Espaciais e Expressões de Posicionamento

O termo “relacionamento espacial” é utilizado para definir relacionamentos que envolvem a posição espacial relativa entre objetos, por exemplo, um objeto A “dentro de” ou “perto de” um objeto B [FuAJ03, Rodr02]. Os relacionamentos espaciais podem ser agrupados em três categorias [Guti94]: topológicos, métricos e direcionais. Os relacionamentos topológicos representam o grau de conectividade entre objetos geográficos. São relacionamentos que descrevem os conceitos de vizinhança, incidência e sobreposição, mantendo-se invariantes ante a transformações como escala e rotação. Os relacionamentos métricos descrevem quantidades mensuráveis de forma quantitativa, indicando se os objetos estão perto ou longe, como, por exemplo, “distância < 100”, e “a 5 km de”. Os relacionamentos direcionais descrevem orientação e ordem. A orientação é usada para descrever direções como as cardinais (ex., “ao sul de”). Já os relacionamentos de ordem indicam a ordem total ou parcial entre os objetos espaciais e são descritos por expressões como “em frente a”, “atrás de” e “acima de” [Free75]. Pullar e Egenhofer [PuEg88] consideram mais um tipo de relacionamento espacial, o *fuzzy*. Os relacionamentos *fuzzy* descrevem relacionamentos de proximidade usando termos imprecisos como “perto de”, “na vizinhança de” e “próximo a”, e são dependentes do tipo do objeto, do seu tamanho, da sua forma e do contexto no qual são usados. Muitos autores consideram relacionamentos de orientação como fazendo parte dos relacionamentos *fuzzy*.

Conforme mostrado por Delboni [Delb05], nas páginas da Web brasileira, os relacionamentos métricos, direcionais e *fuzzy*, são dominantes na descrição de uma localização, sendo que os relacionamentos *fuzzy* são os mais freqüentes. Os relacionamentos topológicos aparecem em número bem menor, em geral como parte de expressões de posicionamento.

Uma *expressão de posicionamento* [Delb05] é definida como uma construção semântica em linguagem natural utilizada para expressar, por meio de um par <relacionamento espacial, ponto de referência>, a posição relativa no espaço de um lugar de interesse. Assim, por exemplo, a expressão <shopping, perto da, UFMG> associa um assunto (shopping) à expressão de posicionamento <perto da UFMG>. O estudo relatado em [DeBL05, Delb05] indica que a partir de expressões de posicionamento, os dez relacionamentos espaciais mais encontrados em um contexto urbano, no Brasil, são: "próximo a", "em frente a", "a N km de", "ao lado de", "perto de", "a N minutos de", "a N quilômetros de", "dentro de", "atrás de" e "nas proximidades de". Esse estudo mostra ainda que expressões que indicam proximidade e expressam distâncias aproximadas são muito comuns na Web brasileira. Através de experimentos, Delboni [Delb05] mostrou que esses relacionamentos e suas variações em linguagem natural podem ser considerados equivalentes em termos da distância que descrevem em relação a um ponto de referência intra-urbano. Por exemplo, dois locais podem ser definidos como estando "próximos" ou "perto" um do outro ou, equivalentemente, a "500 metros" ou "poucos minutos".

Em qualquer raciocínio espacial qualitativo existe uma tolerância de imprecisão, uma vez que o dado preciso e quantitativo nem sempre está disponível. No entanto, apesar da imprecisão, essa equivalência permite que muitas inferências válidas sejam feitas usando as expressões de posicionamento encontradas em páginas da Web. Os resultados apresentados em [Delb05] são usados na ontologia de lugar proposta nesta tese.

2.5 Trabalhos Relacionados

A identificação do contexto geográfico em páginas da Web tem gerado um crescente interesse nos setores comerciais e de pesquisa. Vários trabalhos de pesquisa abordam o reconhecimento e o uso de informações espaciais em páginas da Web, demonstrando ser

possível descobrir e explorar a informação geográfica. O uso dessa informação provê um novo paradigma de navegação e recuperação de informação na Web [DiGS00].

Kambayashi et al. [KaCL01] classificam esses trabalhos em três grandes categorias. A primeira categoria utiliza os mapas como uma interface amigável onde é possível manipular os dados geográficos através de navegadores Web. Essa abordagem é chamada de *map-enhanced Web applications*. A segunda categoria, baseada em conteúdo, associa um "lugar" às páginas da Web usando a informação de localização geográfica encontrada nas páginas. Essa informação consiste em nomes de lugares, coordenadas geográficas e endereços, que são usados para classificar e indexar as páginas. A terceira categoria trata da integração da informação na Web com o conhecimento geográfico, possibilitando novas descobertas. Trabalhos recentes, como Geo-Tumba! e SPIRIT, descritos a seguir, abrangem todas essas categorias.

Os trabalhos descritos nesta seção e que se relacionam com esta tese pertencem, na sua maioria, à segunda e terceira categorias. Para uma melhor compreensão, a descrição desses trabalhos será apresentada segundo três direções de pesquisa identificadas dentro da segunda e terceira categorias: (1) exploração de informação geográfica em páginas da Web, (2) identificação e tratamento de ambigüidades entre nomes de lugar e (3) uso de *gazetteers* como bases de conhecimento. Além disso, são abordados alguns sistemas e serviços de busca geográfica disponíveis na Web e trabalhos relacionados à extração de dados de fontes da Web.

2.5.1 Exploração da Informação Geográfica em Páginas da Web

Segundo Larson [Lars96], Recuperação de Informação Geográfica (RIG) é uma área de pesquisa aplicada que combina aspectos de bancos de dados, interação homem-máquina, sistemas de informação geográficos (SIG) e recuperação de informação (RI).

Em RIG, o reconhecimento de um contexto geográfico em páginas da Web apresenta duas abordagens principais [AHSS04]. A primeira abordagem, chamada de *Geografia da Fonte* (*Source Geography*) é relacionada com a origem das páginas e utiliza os elementos de infra-estrutura da Internet para obtenção de informações sobre a localização física dos servidores onde esses documentos estão hospedados. A segunda abordagem, denominada *Geografia do Alvo* (*Target Geography*), é baseada no conteúdo das páginas, utilizando elementos nelas contidos para deduzir uma ou mais localizações

encontradas em seus textos. Esses elementos consistem em nomes de lugar, coordenadas geográficas, códigos postais e endereços que são usados para classificar e indexar as páginas.

A primeira abordagem é mais limitada do que a segunda, porque não é possível concluir que o conteúdo de uma página está relacionado com o lugar onde ela foi criada. Já a abordagem baseada em conteúdo, apesar de mais rica, encontra como desafios a extração de termos geográficos em textos não estruturados e a identificação e remoção de ambigüidades no processo de atribuição de um nome de lugar. Dentro da primeira abordagem, encontram-se os trabalhos de Buyukkokten et al. [BCGG99] e de Ding et al. [DiGS00]. Este último combina a primeira e a segunda abordagens, considerando não só informações de infra-estrutura como também o conteúdo.

O trabalho de Buyukkokten et al. [BCGG99] foi um dos primeiros a reconhecer e considerar o uso da informação geográfica em *sites* da Web. O artigo introduz o conceito de *sites* de "interesse local" e de "interesse global". Um *site* de "interesse local" é aquele relevante dentro de um determinado limite geográfico ou para usuários próximos dessa localização, como restaurantes, teatros, imóveis de aluguel, etc. Já o de "interesse global" é aquele relevante para a comunidade Web, como jornais *on-line*, instituições bancárias e lojas *on-line*. Um escopo geográfico é atribuído ao *site* baseado na distribuição geográfica das páginas da Web que apontam para ele. Por exemplo, o jornal *New York Times* tem mais reputação global que o *San Francisco Chronicle*, baseado no fato de que ele é mais apontado por um conjunto de *sites* geograficamente diverso. Para a obtenção desse escopo geográfico, o artigo discute como mapear um *site* da Web para a localização geográfica onde ele foi criado e apresenta uma ferramenta que mostra a origem geográfica dos *hyperlinks* que apontam para uma determinada página da Web. A ferramenta ajuda a visualizar o escopo geográfico das páginas da Web. A Figura 2.1 exemplifica o uso dessa ferramenta na visualização do escopo geográfico do *site* do jornal *San Francisco Chronicle*[13]. Cada localização geográfica com *link* para http://www.sfgate.com é representada por um círculo. Como pode ser observado, o jornal *San Francisco Chronicle* é popular principalmente na área metropolitana de São Francisco (*Bay Area*) e na Califórnia, tendo poucos leitores distribuídos em outros estados.

[13] `http://www.sfgate.com`

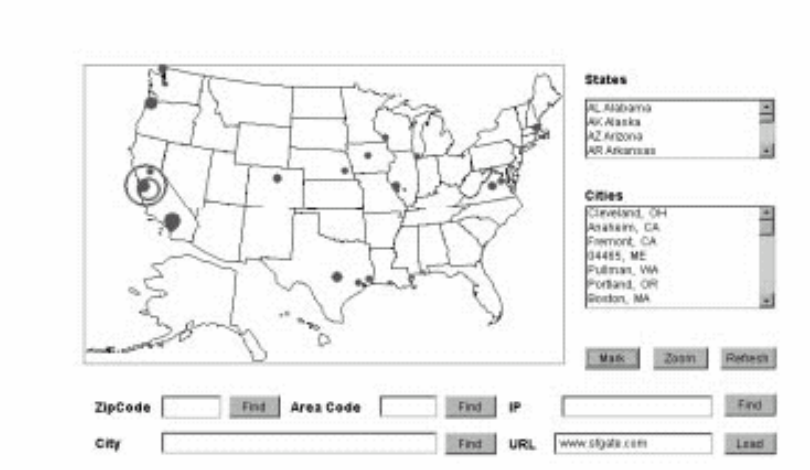


Figura 2.1 – Distribuição geográfica dos *hyperlinks* para http://www.sfgate.com

Ding et al. [DiGS00] estenderam a abordagem apresentada em [BCGG99]. Para computar o escopo geográfico de páginas da Web, eles propuseram duas estratégias complementares: a distribuição geográfica dos *hyperlinks* que apontam para o *site* e a distribuição, no texto da página, das referências geográficas (nomes de locais). O escopo geográfico é definido como a área geográfica que o criador do *site* pretende atingir (Figura 2.2). O artigo apresenta um protótipo de uma máquina de busca (*GeoSearch*) que descarta qualquer página que não esteja no escopo geográfico do usuário (Figura 2.3). No entanto, essa técnica não determina se a consulta seria mais bem atendida com páginas globais ou locais.

Figura 2.2 – Escopo geográfico dos jornais New York Times, The Digital Missourian e *The Stanford Daily*

Figura 2.3 – Resultados de uma busca no protótipo da máquina de busca Geosearch

Assim como Ding et al. [DiGS00], o trabalho apresentado por McCurley [Mccu01] abrange as abordagens de conteúdo e de infra-estrutura. McCurley foi o primeiro a apresentar uma variedade de abordagens para reconhecimento de referências geográficas em páginas da Web. Ele analisou vários aspectos das páginas da Web que poderiam ter uma associação geográfica, como sua URL, o idioma em que a página foi escrita e a existência de um número de telefone ou código postal. McCurley propõe o uso de *gazetteers* para o reconhecimento de nomes de entidades geográficas. Para páginas que não possuem um contexto geográfico definido, propõe o uso de *hyperlinks*. Esta é uma maneira de verificar se as páginas apontadas pelas páginas que não possuem contexto geográfico, ou as páginas que apontam para as páginas sem contexto geográfico, possuem uma associação geográfica definida que poderia ser usada para atribuir o contexto geográfico. McCurley aborda também o problema de indexação e apresenta um sistema experimental para navegação geográfica na Web, onde as páginas são apresentadas como pontos em um mapa (Figuras 2.4 e 2.5). A função "*where*" localiza as páginas no mapa (Figura 2.4) e a função "*what's nearby*" mostra uma visão do mapa que contém as páginas que, pelo contexto geográfico, estão próximas da página que está sendo visualizada.

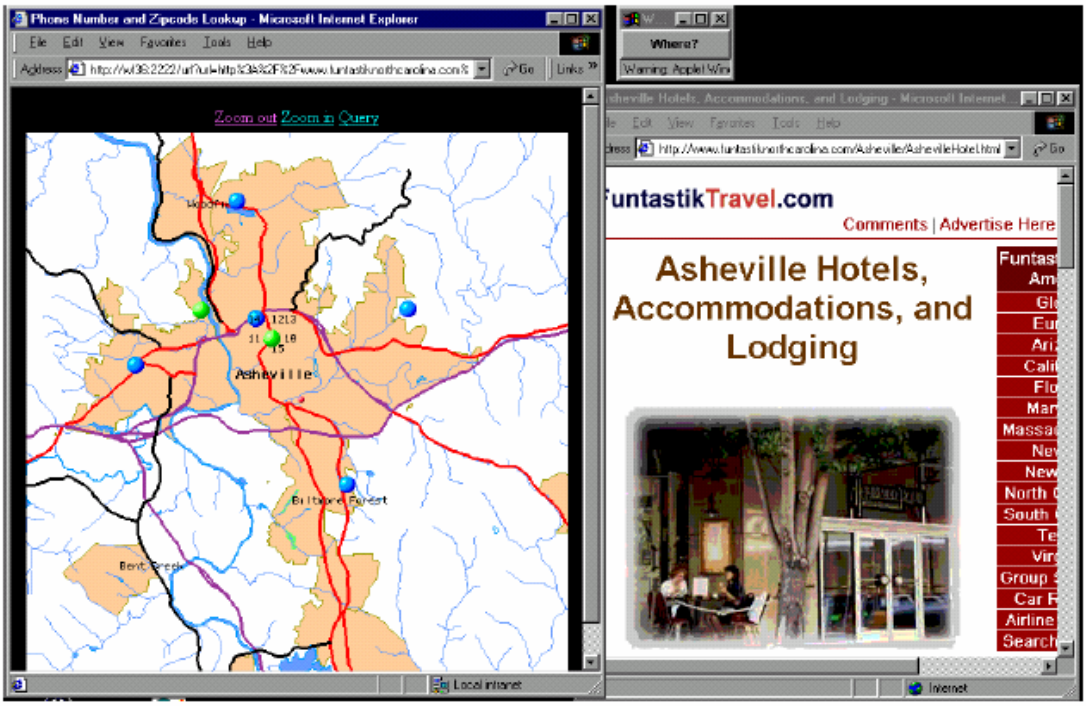


Figura 2.4- Visualização da página do Hotel Asheville, após uso da opção "*where*", que abre uma nova janela mostrando no mapa a cidade de Asheville

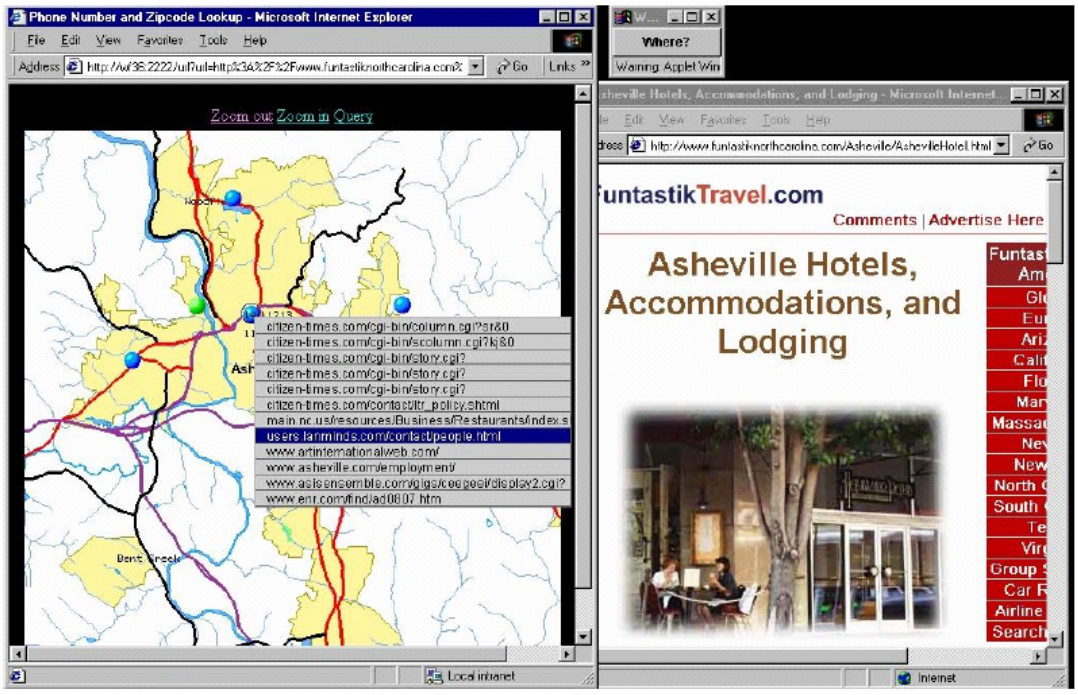


Figura 2.5 –Visualização de uma lista de URLs associadas a uma localização no mapa. Com a escolha de uma URL, o navegador abre uma nova janela com a página original, no caso a página do Hotel Asheville

O trabalho de Wang et al. [WXWL05] propõe um sistema que emprega um conjunto de algoritmos e diferentes fontes geográficas para extrair informação geográfica a partir de conteúdo da Web, mineração da estrutura de *hyperlinks* e *logs* do usuário. O sistema considera três tipos de localização na Web: (1) localização do provedor (*provider location*), (2) localização do conteúdo (*content location*) e localização do escopo geográfico que o *site* atinge (*serving location*). O conjunto de algoritmos desenvolvidos para computar os diferentes tipos de localização na Web é baseado nas suas características específicas, incluindo o algoritmo *top-down* proposto para estimar a localização dominante no conteúdo. Os autores estenderam o algoritmo proposto por Ding et al. [DiGS00], incluindo fontes geográficas mais confiáveis como códigos postais, números de telefones e localizações de usuários Web. Os resultados experimentais mostraram que a abordagem proposta obteve resultados melhores que o algoritmo genérico de detecção de lugar.

Outros trabalhos usam a Web como uma base de conhecimento, onde novas descobertas podem ser obtidas através de inferências e do raciocínio espacial. Entre eles estão os de Rodríguez [Rodr02] e Morimoto et al. [MAHM03].

Rodríguez [Rodr02] apresenta uma abordagem baseada no conhecimento geográfico (*knowledge-based approach*) para aumentar a capacidade de resposta das máquinas de busca atuais em consultas expressas por operadores espaciais tais como “em”, “norte”, “oeste”, “perto”. Como exemplo, é citada a consulta “encontre hotéis em Pucón, Chile, e nas cidades adjacentes”. Para atender a esse tipo de consulta é descrito um modelo para organização e derivação de relacionamentos espaciais com base na estrutura hierárquica do espaço (organizado em regiões) e nas inter-relações das regiões conectadas. A idéia geral é mapear o nome de regiões e lugares dentro dessa estrutura semântica, de tal modo que as páginas da Web associadas a esses lugares não necessitem incorporar propriedades geométricas. As análises envolvendo os relacionamentos topológicos, cardinais e de distância são favorecidas ao se utilizar o raciocínio espacial. O trabalho apresenta diretrizes de como esse modelo pode ser usado para estender as técnicas atuais de busca a páginas da Web. No entanto, não aborda como mapear um nome de lugar para a estrutura semântica do espaço proposta.

O trabalho de Morimoto et al. [MAHM03] apresenta o problema de extração de conhecimento espacial a partir do conteúdo da Web. Eles partiram do princípio que

páginas comerciais na Web normalmente contêm endereço, telefone e também descrições dos produtos e serviços oferecidos. Assim, seria natural supor a existência de uma associação entre o conteúdo de uma página da Web e alguma informação de localização específica contida nela. Para verificar essa hipótese, eles apresentaram um sistema que extrai informação geográfica de localização e termos encontrados em páginas da Web. Para cada endereço postal encontrado (G1 na Figura 2.6), são aplicadas técnicas de geocodificação para a atribuição de coordenadas geográficas (latitude/longitude) (G2 na Figura 2.6). Simultaneamente são extraídos conceitos da coleção de páginas - palavras-chave mais significativas (C1 na Figura 2.6). São eliminados os marcadores HTML e extraídos nomes, termos, abreviaturas e outros termos significativos. A coleção de páginas é representada através do modelo de espaço vetorial [BYRN99] . Cada página é representada por um vetor consistindo em um conjunto de palavras-chave (termos) com peso. Para cada conjunto de termos extraídos é criada uma matriz tf-idf (*term frequency* x *inverse document frequency*) onde cada vetor corresponde a uma página individual e cada atributo de um vetor corresponde a um termo do conjunto de termos extraídos (C2 na Figura 2.6). O próximo passo é a redução do número de colunas da matrix tf-idf para diminuição do custo computacional. Depois os termos são agrupados e rotulados com várias palavras-chave que indicam conceitos associados com o conteúdo da página (C5 na Figura 2.6).

Uma tabela de associação geo-espacial é criada contendo o identificador da página, as coordenadas geográficas e os conceitos. Aplicando técnicas de mineração de dados geo-espaciais a esta tabela e uma função que encontra conjuntos de classes cujos objetos estão espacialmente perto uns dos outros, é possível encontrar padrões espaciais de comportamento. Os resultados experimentais apresentados para um conjunto de páginas do Vale do Silício na Califórnia, mostraram, por exemplo, que páginas de determinados conteúdos como, *softwares* e telefone, estavam fisicamente localizadas muito próximas umas das outras. No entanto, falsos positivos apareceram na tabela de associação provando ser necessário um trabalho de refinamento da idéia, além de um método de *geoparsing* com mais acurácia e eficiência.

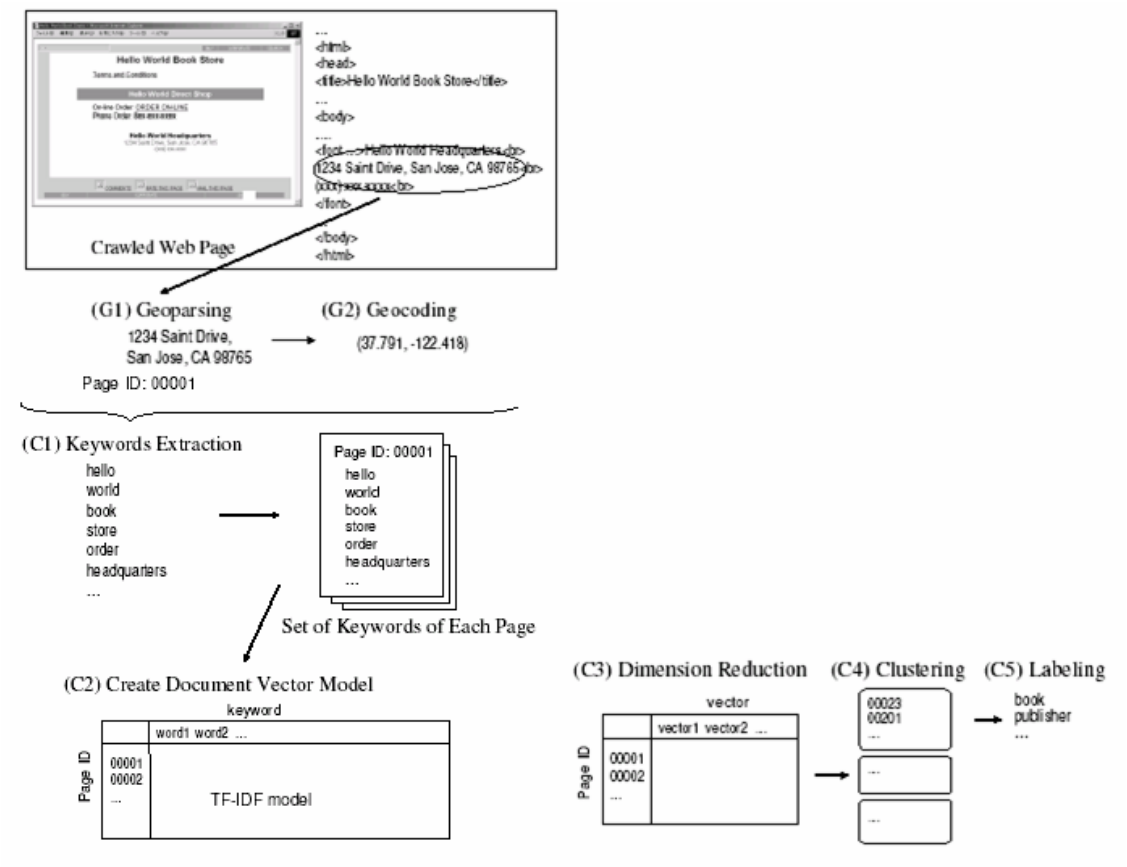


Figura 2.6 – Extração da informação geográfica

2.5.2 Identificação e Tratamento de Ambigüidades entre Nomes de Lugar

Na abordagem baseada em conteúdo, a tarefa de se associar um lugar às páginas da Web é dividida em três sub-problemas: (a) extração dos nomes de lugar, (b) eliminação de ambigüidades e (c) associação de um lugar às páginas.

A extração de nomes de lugar de página da Web é um problema relacionado à tarefa de extrair e distinguir o nome de diferentes tipos de entidade em texto, como pessoas, organizações, eventos ou entidades geográficas. Este é um assunto muito explorado em processamento de linguagem natural, sendo conhecido como *Named Entity Recognition* (NER) [Chin97]. O reconhecimento de nomes de lugares (entidades geográficas) é uma sub-categoria do problema de reconhecimento de nomes de entidades e recentemente tem chamado a atenção da comunidade de recuperação de informação geográfica na Web. A extração de nomes de lugar possui algumas peculiaridades que a diferenciam da extração do nome de outras entidades. Uma dessas peculiaridades é a questão da ambigüidade. A maioria dos nomes de lugar encontrada na Web é ambígua [AHSS04].

Da mesma forma que um mesmo nome identifica vários lugares, como Alagoinha no Ceará, em Pernambuco e na Paraíba, um lugar pode ser referenciado por vários nomes diferentes, como Belo Horizonte, BH e Belô. Uma outra peculiaridade é o próprio reconhecimento do nome de um lugar, uma vez que vários desses nomes são homônimos de nomes de pessoas como, por exemplo, no caso das cidades de Adhemar de Barros (PR) e Afonso Claudio (ES).

Amitay et al. [AHSS04] preocuparam em associar um local a uma página como um todo – o que eles denominam de *foco geográfico*. O sistema Web-a-Where descrito em [AHSS04] é baseado em um *gazetteer* que associa uma localização geográfica ao conteúdo da página (*foco geográfico*). As ambigüidades são classificadas em dois tipos: *geo/geo* (ambigüidade entre dois nomes de lugares como Rio de Janeiro, Brasil, e Rio de Janeiro, Peru) e *geo/não-geo* (ambigüidade entre um nome de lugar e um nome que não é um lugar, como Tiradentes, Minas Gerais, e Tiradentes, o inconfidente). São usadas várias técnicas para resolver esse problema de ambigüidade que utilizam, entre outros, população e valores de confiança derivados de qualificadores encontrados na vizinhança (por exemplo, nome da cidade imediatamente seguido pelo nome do estado). A estratégia proposta para se obter o foco geográfico da página seleciona até quatro nomes de lugar que cobrem geograficamente a maior parte dos nomes contidos na página. Para isso é usado o relacionamento hierárquico entre esses lugares encontrado em um *gazetteer*.

Já em [ZWSL05], um nome de lugar é associado a uma página da Web como um todo ou a partes dela, quando nela existir um conjunto de lugares. Para extrair nomes de lugar dentro dos Estados Unidos, os autores modificaram o *software* GATE[14], desenvolvido para extrair nomes de entidades, incorporando o *gazetteer* US Census 2000 [USCB00]. O trabalho propõe regras para associação de um lugar a segmentos de uma página e a toda a página. Essas regras são baseadas em dois métodos: um remove ambigüidade entre nomes e outro associa um lugar à página. As ambigüidades são tratadas com um conjunto de regras que consideram informação de contexto das páginas e as distâncias euclidianas entre os lugares cujos nomes estão sendo analisados. O método usado para a associação de lugar incorpora a hierarquia de lugares encontrada no *gazetteer*, de forma a associar o nome de lugar mais apropriado ao lugar encontrado.

[14] `http://www.gate.ac.uk`

O trabalho mostra ainda que o método usado para remover ambigüidades entre nomes de lugar funciona bem para ambigüidades do tipo *geo/geo*. Além disso, segundo os autores, ao se associar um lugar às páginas da Web, elas poderão ser acessadas tanto pelo nome do lugar associado quanto por sua localização geográfica.

Como visto, o uso de *gazetters* no tratamento de ambigüidades é uma técnica comum que foi validada por Mikheev et al. [MiMG99]. Esses autores tentaram evitar o uso de *gazetteers* no reconhecimento de nomes de lugar, concluindo que a extração de nomes de lugar sem o uso de um *gazetteer* produz resultados muito ruins (precisão/revocação de 46% / 59% respectivamente).

Schilder et al. [ScVH04] usam expressões espaciais para a remoção de ambigüidades entre nomes de lugar. O foco do trabalho é a extração e derivação do significado de expressões espaciais, sua classificação quanto ao uso e associação com descrições de um evento. O método foi aplicado a artigos de jornais coletados de *sites* da Web. Os autores procuram associar entidades geográficas e geopolíticas ao evento encontrado, concluindo que a acurácia de um sistema pode melhorar se diferentes técnicas para eliminar ambigüidades forem usadas.

Poucos são os trabalhos envolvendo o uso de relacionamentos espaciais no reconhecimento de nomes de lugar em páginas da Web, na especificação de consultas geográficas e na atribuição do contexto geográfico. Os trabalhos de Silva et al. [SMCC06], Heinzle et al. [HeKS03], Sanderson e Kohler [SaKo04], e Egenhofer [Egen02] destacam a necessidade de as máquinas de busca entenderem e fazerem uso de relações espaciais como forma de reconhecer consultas geográficas e melhorar a qualidade das respostas a essas consultas.

2.5.3 Uso de *Gazetteers* como Bases de Conhecimento

Para extração de nomes de lugar, alguns trabalhos optaram por construir uma base de conhecimento sobre lugares estruturada de acordo com uma ontologia, ao invés de fazer uso de algum *gazetteer* já existente. Essa abordagem apresenta várias vantagens em relação ao uso de um *gazetteer* tradicional porque são adicionados novos relacionamentos, permitindo dividir o espaço em mais níveis hierárquicos e possibilitando o reconhecimento de entidades geográficas no nível correto de "granularidade" de acordo com o objetivo da aplicação. Por exemplo, o reconhecimento

pode ser feito no nível de um país, cidade, bairro ou local específico associado a um endereço. Os trabalhos de Manov et al. [MKPB03], Chaves et al. [ChSM05] e Fu et al. [FuAJ03, FuJA05] utilizam esta abordagem.

Manov et al. [MKPB03] criaram uma base de conhecimento de locais com o objetivo de ajudar no processo de extração de informação dentro do projeto KIM (*Knowledge Information Management*). A base de conhecimento de locais foi estruturada de acordo com a sub-ontologia *Location*, que é parte da KIMO (KIM *Ontology*). A base de conhecimento foi criada somente com locais considerados mais importantes, como continentes, regiões globais, países, cidades grandes, rios, montanhas, oceanos e monumentos, entre outros, contendo aproximadamente 50.000 locais. Cada local possui seu nome em inglês, francês e, algumas vezes, no idioma local. A eliminação de ambigüidades entre nomes de lugar é executada usando padrões baseados em gramáticas de linguagem natural.

Chaves et al. [ChSM05] criaram uma base de conhecimento geográfico (*Geographic Knowledge Base* - GKB) que integra dados e conhecimentos obtidos de várias fontes em um único esquema. GKB faz parte do projeto GREASE - *Geographic Reasoning for Search Engines*[15], que pesquisa métodos, algoritmos e arquiteturas de *software* para atribuir escopo geográfico a páginas da Web e recuperar documentos usando feições geográficas. GKB contém informação de entidades geopolíticas e geofísicas de Portugal, como também dos recursos de rede como *sites* da Web e domínios da Internet. GKB também possui informações sobre as principais regiões do mundo em quatro idiomas e usa uma ferramenta para geração de ontologias de nomes geográficos, segundo o padrão OWL.

Fu et al. [FuAJ03, FuJA05] descrevem uma ontologia de lugar que modela tanto o vocabulário quanto a estrutura espacial de lugares, com o propósito de possibilitar a recuperação de informação geográfica na Web. Essa ontologia é o componente mais importante da arquitetura do projeto SPIRIT – *Spatially-Aware Information Retrieval on the Internet*[16]. No SPIRIT, a ontologia de lugar é usada para ajudar na eliminação de ambigüidades nas consultas, na expansão dos termos da consulta e na ordenação dos resultados [FuJA05, JPRS02]. A ontologia de lugar é representada na linguagem

[15] `http://xldb.di.fc.ul.pt/grease`
[16] `http://geo-spirit.org`

DAML+OIL [HoHa00] e foi inicialmente criada com lugares do Reino Unido. Os experimentos demonstraram a importância de se combinar o casamento do nome com a posição na hierarquia espacial para verificar a similaridade entre lugares [FuJA05].

2.5.4 Sistemas de Busca Geográfica na Web

O reconhecimento do potencial da busca geográfica na Web tem crescido nos últimos anos. Uma pesquisa realizada pelo *The Kelsey Group*[17] e o *BizRate.com*, em janeiro de 2004, revelou que 25% das buscas de compradores *on-line* eram por produtos ou serviços próximos às suas casas ou locais de trabalho. Com isto, têm surgido propostas de máquinas de busca com capacidade de execução de consultas geográficas (Geo-Tumba!, SPIRIT), serviços específicos para recuperação de informação geográfica nas máquinas de busca convencionais (*Yahoo! Local, Google Local, AOL Local Search*) e também sistemas comerciais (*Geosearch*, *MetaCarta*, *InfoSpace*).

MetaCarta[18] é um sistema comercial de busca, baseado na geografia, que faz a extração de nomes de lugares, remove ambigüidades entre esses nomes e executa o processamento de consultas baseado no nome de lugar. Um *gazetteer* é usado na identificação de nomes de lugar em páginas da Web. Cada nome de lugar no *gazetteer* é associado a um valor de confiança, determinado pela probabilidade dele ter sido corretamente reconhecido. Padrões de processamento de linguagem natural, convenções de uso de letra maiúscula, nomes de lugar encontrados na vizinhança, população dos lugares e outras heurísticas são utilizadas no processo de remoção de ambigüidades. *MetaCarta* também processa referências geográficas relativas como "15 milhas noroeste de Portland" e é apropriado para busca de páginas que possuem conteúdo geográfico que não seja determinado por endereço postal ou número de telefone [RaBB03] .

InfoSpace[19] é um sistema comercial que, entre outras funções, localiza serviços já previamente cadastrados por tipo, nome, endereço e raio de proximidade. A busca por proximidade, permite encontrar serviços como, restaurantes, hotéis, aeroportos, bancos, perto de uma determinada localização retornando uma lista de serviços com a respectiva distância. No entanto, a visualização só é feita após a escolha de um dos serviços da lista.

[17] http://www.kelseygroup.com/press/pr040211.htm
[18] http://www.metacarta.com
[19] http://www.infospace.com/home/yellow-pages/near-address

Geosearch foi uma das primeiras máquinas de busca especializadas em busca local (*local search*) e esteve disponível na Web de abril de 2000 a março de 2002. Propriedade da Vicinity, foi implementada em associação com a Northern Light Search Engine[20], tendo sido o primeiro esforço em larga escala para derivar conteúdo local diretamente da Web. A idéia básica do *Geosearch* era transformar a informação de localização contida nas páginas da Web para uma forma que as máquinas de busca pudessem usar para efetuar uma busca eficiente de proximidade. Seus desenvolvedores apostaram na idéia de que uma porção significativa do conteúdo das páginas da Web relevante para a busca local continha um endereço postal bem formado ou um número de telefone ou os dois. O módulo reconhecedor de endereços detectava endereços no formato norte-americano e canadense e números de telefone dos Estados Unidos, Canadá e Caribe. Após encontrado, o endereço era geocodificado ou seja, coordenadas geográficas lhe eram atribuídas. Segundo Himmelstein [Himm05], o *Geosearch* excedeu as expectativas, como uma prova de conceito de larga escala, demonstrando que a Web é uma rica fonte de conteúdo local.

Google Local[21], *Yahoo! Local*[22] e o *AOL Local Search*[23] são opções das respectivas máquinas de busca para a localização de serviços. Entretanto, a técnica empregada e detalhes funcionais desses serviços não foram publicados. Apesar de o *Google Local* utilizar praticamente a mesma abordagem do *Geosearch*, ele acrescentou uma correlação entre a informação de endereço encontrada nas páginas da Web e os dados das páginas amarelas na Web. O *Google Local* localiza lojas e serviços na vizinhança de uma localização geográfica específica. Essa abordagem permite que o usuário especifique através de um raio qual a área onde a consulta irá ser processada. Atualmente, a consulta está restrita aos Estados Unidos e Canadá, e limitada aos serviços contidos nas páginas amarelas. Na consulta de proximidade, o resultado é ordenado pela página mais relevante dentro de uma determinada distância. Não existe a opção de a ordenação ser pelas páginas relativas à região mais próxima, que contenham o serviço desejado e que estejam contidas no intervalo de distância solicitado. Ainda não é possível também fazer consultas combinadas, por exemplo, "hotel e restaurante a 5 milhas de um local informado".

[20] `http://northernlight.com`
[21] `http://local.google.com`
[22] `http://local.yahoo.com`
[23] `http://localsearch.aol.com`

Já o trabalho de Hiramatsu e Ishida [HiIs01], também na linha da busca local (*local search*), propôs uma nova abordagem para integrar o ambiente Web e Sistemas de Informação Geográficos, combinando as URL das páginas de serviços com as funções de um SIG. O objetivo é refletir, num mapa, a *Web local* de forma a ajudar as pessoas de áreas metropolitanas a localizar serviços utilizados no seu dia-a-dia. Quando um serviço é selecionado no mapa da cidade, o sistema mostra a página Web associada àquele serviço.

SPIRIT[24] (*Spatially-Aware Information Retrieval on the Internet*) é um projeto patrocinado pelo *European Commission Framework*, cujo principal objetivo é criar ferramentas e técnicas para ajudar as pessoas a encontrar na Web informações relacionadas com uma localização geográfica específica [JPRS02]. O SPIRIT encontra-se em fase de desenvolvimento e pretende atender a necessidade de dois grupos de usuários: os que desejam localizar informações sobre um tópico de interesse ou fenômeno que ocorre ou está associado a algum lugar no espaço (como serviços: lojas, cinemas ou restaurantes) e os usuários de sistemas de informação geográficos que estão interessados em algum tipo de informação geo-espacial em meio digital (como mapas, imagens ou modelos de terrenos). O projeto inclui o desenvolvimento de uma máquina de busca com recursos espaciais que permita consultas via palavras-chave, via mapa e via um croqui da área desejada. Além disso, estão sendo pesquisados diferentes métodos que combinem o índice textual e geográfico. Vaid et al. [VJJS05] descrevem três métodos que combinam a indexação espacial e textual. O índice espacial se mostrou competitivo, em termos de velocidade e desempenho de armazenamento, em comparação experimental com um índice convencional usado em máquinas de busca.

Silva et al. [SMCC06] descrevem métodos para determinar o escopo geográfico de páginas da Web na máquina de busca portuguesa Tumba! - Temos Um Motor de Busca Alternativo[25]. O objetivo é o de melhorar a qualidade das respostas às consultas geográficas (*Geo-Tumba Geographical Web Search Engine*). As páginas coletadas são transformadas em documentos XML/RDF 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 (GKB) [ChSM05] para definir o escopo geográfico das páginas. A atribuição de escopo geográfico é feita por um

[24] `http://geo-spirit.org`
[25] `http://www.tumba.pt`

método que usa o reconhecimento de nomes de entidades geográficas junto com um algoritmo de "combinação/eliminação de ambigüidades" construído sobre os relacionamentos da ontologia. Estão sendo estudadas heurísticas para resolver ambigüidades semânticas dos termos geográficos com o objetivo de tornar essa fase mais eficaz. É feita também uma análise de *links* entre as páginas, considerando o escopo geográfico. Por exemplo, 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*). 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. *Geo-Tumba!* faz parte do projeto GREASE.

2.5.5 Extração de Dados de Páginas da Web

Nos últimos anos, vários trabalhos encontrados na literatura têm abordado o problema de extração de dados de páginas da Web. Ferramentas para extração desses dados continuam sendo propostas, visando a geração automática ou semi-automática de *wrappers* que permitam extrair informações mais precisas e que estejam de acordo com a intenção do usuário.

As abordagens presentes na literatura que tratam desse problema usam técnicas originárias de diversas áreas, como, por exemplo, processamento de linguagem natural, linguagens e gramáticas, recuperação de informação, *machine learning*, bancos de dados e ontologias. Em Laender et al. [LRST02] foi introduzida uma taxonomia para classificação de ferramentas de extração de dados de acordo com o tipo de técnica empregada para a geração dos *wrappers*. Foram definidas seis categorias: (1) linguagens para geração de *wrappers*, (2) ferramentas baseadas na estrutura HTML, (3) ferramentas baseadas em processamento de linguagem natural, (4) ferramentas baseadas em indução, (5) ferramentas baseadas em modelagem de dados e (6) ferramentas baseadas ontologias. As cinco primeiras categorias adotam abordagens sensíveis ao contexto enquanto que a sexta categoria considera apenas o conteúdo. A seguir, baseado em [LRST02], é descrita sucintamente cada uma das categorias mencionadas acima.

Linguagens para a geração de *wrappers*. Essas linguagens oferecem facilidades específicas para auxiliar na tarefa de geração de *wrappers* constituindo uma abordagem eficiente para extração de dados de páginas da Web. Exemplos de ferramentas que adotam essa abordagem são: TSIMMIS [HaMG97], Minerva [CrMe98] e Web-OQL [ArMe98].

Ferramentas baseadas na estrutura HTML. As ferramentas baseadas na estrutura HTML utilizam as características estruturais inerentes das páginas HTML, transformando as páginas em uma árvore que representa a hierarquia de seus marcadores. Regras de extração são geradas semi- e automaticamente para extrair dados nas raízes da árvore. Essa abordagem pressupõe o uso consistente de marcadores e estruturas HTML (tabelas, listas, etc.), propiciando geralmente um bom grau de automação. Algumas ferramentas representativas dessa abordagem são: W4F [SaAz00], XWRAP [LiPH00], RoadRunner [CrMM01] e Lixto [BaFG01].

Ferramentas baseadas em processamento de linguagem natural. As ferramentas dessa categoria utilizam basicamente técnicas de processamento de linguagem natural (PLN) para gerar regras de extração com base em restrições sintáticas e semânticas existentes nos documentos. Usualmente requerem um pré-processamento para identificação de elementos sintáticos no documento. Essa abordagem é mais adequada para a extração de dados em páginas da Web que contenham texto livre como, por exemplo, notícias, classificados e anúncios. Ferramentas representativas dessa categoria são: WHISK [Sode99], RAPIER [CaMo99] e SRV [Frei00].

Ferramentas baseadas em indução. Essas ferramentas utilizam técnicas de *machine learning* para geração semi-automática de *wrappers*. Recebem como entrada um conjunto de exemplos do que deve ser extraído gerando como saída regras de extração baseadas nos separadores que delimitam os dados de interesse. Isso faz com que essas ferramentas sejam mais convenientes para documentos HTML do que as ferramentas baseada em PLN. Exemplos dessas ferramentas são: SoftMealy [HsDu98], WEIN [Kush00] e STALKER [MuMK01].

Ferramentas baseadas em modelagem. As ferramentas baseadas em modelagem requerem uma descrição da estrutura dos objetos (dados) a serem extraídos. Também utilizam exemplos para geração de padrões de extração e procuram encontrar nas

páginas-alvo, objetos com estrutura e contexto semelhantes aos dos exemplos. Usam técnicas semelhantes às utilizadas pelas ferramentas de indução. Exemplos de ferramentas que adotam essa abordagem são: NoDoSE [Adel98] e DEByE [LaRS02].

Ferramentas baseadas em ontologias. As ferramentas baseadas em ontologias requerem a construção prévia de uma ontologia para o domínio de aplicação considerado. O processo de extração consiste em encontrar nas páginas-alvo cadeias de caracteres que satisfazem as propriedades descritas na ontologia. Uma vez construída, a ontologia é aplicável a qualquer página do domínio considerado. A principal vantagem dessa abordagem em relação às demais é o fato de que, quando as páginas da Web são alteradas, as regras de extração continuam válidas. A ontologia só é alterada quando os requisitos do domínio são modificados. A ferramenta mais representativa dessa abordagem é a desenvolvida pelo Brigham Young University Data Extraction Group (BYU) [ECJL99, Embl04]. A estratégia de extração de dados proposta por esse grupo, utiliza ontologias como base semântica para a extração e estruturação de dados semi-estruturados. As ontologias são previamente definidas visando descrever domínios específicos de interesse, incluindo relacionamentos, aparência léxica e palavras-chave indicadoras de contexto. Usando esse tipo de ontologia, chamada de *ontologia de extração*, uma ferramenta de extração pode automaticamente carregar um banco de dados através do reconhecimento e extração de dados presentes nas páginas da Web.

Esta tese segue a abordagem baseada em ontologias para o reconhecimento e a extração de evidências geo-espaciais presentes em páginas da Web. A ontologia de extração utilizada é a ontologia de lugar urbano *OnLocus* descrita em detalhes no Capítulo 3.

2.5.6 Resumo dos Trabalhos

As Tabelas 2.1, 2.2, 2.3, 2.4 e 2.5 resumem os principais pontos abordados nos trabalhos relacionados.

Tabela 2.1 – Exploração da informação geográfica

Referência	Tipo de Abordagem Utilizada	Tratamento de Ambigüidade	Uso de *Gazetteer*	Tipo de Abordagem para Associação de Lugar
BCGG99	infra-estrutura	Não	Não	Escopo geográfico - *Sites* de interesse local e global
DiGS00	infra-estrutura e conteúdo	Sim	Sim	Escopo geográfico - a área que o criador da página pretende atingir
Mccu01	infra-estrutura e conteúdo	Menciona sem entrar em detalhes	Sim	Contexto geográfico - variedade de abordagens para reconhecimento de lugar nas páginas.
WXWL05	infra-estrutura e conteúdo	Sim. Não entra em detalhes	Sim	Lugar é associado de acordo a localização do provedor (*provider location*), a localização baseada no conteúdo (*content location*) e a localização da área atingida (*serving location*).
Rodr02	Conhecimento geográfico e raciocínio espacial	Não	Não	
MaHM03	Conteúdo	Não	Não	Associa coordenadas geográficas a cada lugar encontrado.

Tabela 2.2 – *Gazetteers* como base de conhecimento

Referência	Criação de uma Ontologia	Conteúdo do *Gazetteer*	Objetivo de Uso
MKPB03	Sim. Uma ontologia de lugar - *Location Knowledge Base*.	Localizações mais importantes: continentes, regiões globais, grandes cidades, rios, oceanos, montanhas e monumentos.	Auxiliar na extração de nomes de lugar.
ChSM05	Sim. *Geographic Knowledge Base* - GKB.	Entidades geopolíticas e geofísicas de Portugal, *sites* Web e domínios de Internet de Portugal. Principais regiões do mundo em quatro idiomas.	Auxiliar na extração de nome de lugar.
FuJA05	Sim. Uma ontologia de lugar.	Inicialmente com lugares do Reino Unido	Auxiliar na recuperação de informação geográfica pela máquina de busca geográfica do Projeto SPIRIT.

Tabela 2.3 – Identificação e tratamento de ambigüidades

Referência	Abordagem Baseada em conteúdo	Tratamento de Ambigüidades	Tipo de Abordagem para Associação de Lugar
ZWSL05	Sim	Usa *gazetteer* e um conjunto de regras. O método proposto funciona bem para ambigüidades geo/geo	Um nome de lugar é atribuído a uma página ou a partes dela.
AHSS04	Sim	Usa *gazetteers* e várias técnicas para resolver ambigüidades. Introduziu os termos ambigüidade geo/geo e geo/não-geo	Foco Geográfico- Atribui um nome de lugar à página como um todo
ScVH04	Sim	Usa expressões espaciais	Atribui um nome de lugar a um evento encontrado em jornais *on line*.

Tabela 2.4– Sistemas de busca geográfica

Produto	**Tipo de Busca**	**Objetivo**
MetaCarta	Busca termos geográficos em textos não estruturados	Recuperar informação geográfica e localizar documentos em mapas. Efetua *geoparsing* em textos não estruturados para encontrar referências a localizações e atribuir coordenadas geográficas
InfoSpace	Busca local	Localiza serviços cadastrados (EUA). Pesquisa por endereço, serviço e distância (raio de proximidade)
Geosearch (extinto)	Busca local	Localizar serviços encontrados em páginas da Web (EUA e Canadá). Pesquisa o que, onde e a qual distância, utilizando na pesquisa endereço e telefone
Google Local	Busca local máquina de busca Google	Localizar serviços (EUA) cadastrados como páginas amarela, associando o *link* da página. Pesquisa por endereço, proximidade e por serviços. A distância não é escolhida. O resultado vem ordenado por distância.
Yahoo! Local	Busca local máquina de busca Yahoo	Localizar serviços cadastrados (EUA e Canadá) .
AOL Local Search	Busca local máquina de busca AOL	Localizar serviços cadastrados, eventos e cinemas (EUA) Pesquisa por endereço e proximidade (*near*). Ordena o resultado
SPIRIT	Máquina de busca com capacidade geográfica	Desenvolver ferramentas e técnicas para auxiliar usuários a encontrar na Web informações relacionadas com uma localização geográfica específica. Projeto Europeu.
GeoTumba!	Máquina de busca com capacidade geográfica	Associar um escopo geográfico a documentos da Web para auxiliar a máquina de busca portuguesa Tumba! a executar pesquisas geograficamente relacionadas. Projeto GREASE

Tabela 2.5– Extração de dados de páginas da Web

Categoria das Ferramentas de Extração	**Características**
Linguagens para geração de *wrappers*	Oferecem facilidades específicas para auxiliar na tarefa de geração de *wrappers*
Ferramentas baseadas na estrutura HTML	Utilizam-se das características estruturais inerentes das páginas HTML
Ferramentas baseadas em PLN	Utilizam técnicas de processamento de linguagem natural (PLN) para gerar regras de extração
Ferramentas baseadas em indução	Utilizam técnicas de *machine learning* para geração semi-automática de *wrappers*
Ferramentas baseadas em modelagem	Requerem uma descrição da estrutura dos objetos (dados) a serem extraídos
Ferramentas baseadas em ontologias	Requerem a construção prévia de uma ontologia para o domínio de aplicação considerado

2.6 Conclusões

Este capítulo apresentou alguns conceitos básicos relacionados ao conteúdo desta tese e discutiu alguns trabalhos relacionados, comerciais e acadêmicos. Apesar de a tese não propor a criação de uma máquina de busca com capacidade geográfica, os trabalhos relacionados ao projeto SPIRIT e ao Geo-Tumba! aproximam-se muito do nosso. No entanto, nossa abordagem se diferencia no objetivo e no grau de granularidade geográfica usado na ontologia de lugar proposta. Como mencionado no Capítulo 1, o trabalho desta tese está focado na *Web local*, reconhecendo e geocodificando localizações intra-urbanas sob forma de endereços mencionados em páginas da Web. Já o Geosearch possui uma abordagem muito parecida com a nossa, entretanto, nosso trabalho propõe estratégias de reconhecimento e extração de endereços baseadas em uma ontologia de lugar e nas características qualitativas e quantitativas dos endereços presentes na Web. Outro aspecto que distingue nosso trabalho dos demais é a questão do tratamento de ambigüidades. Como o nosso reconhecimento de lugar está intimamente ligado a um endereço ou alguma forma indireta de endereçamento, evitamos, assim, o problema de ambigüidade. Nosso método considera um conjunto mínimo de termos que caracterizam unicamente um lugar.

Capítulo 3

"O espaço é a precondição de tudo o que existe,
quer seja na forma material, ou imaterial,
porque nunca podemos imaginar
um objeto ou um ser sem espaço".
Lama Anagarika Govinda

OnLocus: Uma Ontologia de Lugar Urbano

Como visto no Capítulo 2, vários trabalhos utilizam ontologias e *gazetteers* para identificar locais em páginas da Web. Este capítulo propõe uma ontologia de lugar, denominada *OnLocus*, que fornece uma estrutura hierárquica e semântica de localização no espaço geográfico urbano e suas inter-relações. Seu objetivo é auxiliar a extração de contexto geográfico em páginas da Web, possibilitando o reconhecimento e a interpretação de termos referentes a lugares. A ontologia permite associar uma localização a cada evidência geo-espacial encontrada e validada. Essa localização se dá não só em termos de coordenadas geográficas, mas também referenciando a estrutura geopolítica e político-administrativa de um país. Assim, um lugar possui coordenadas geográficas e está situado dentro da hierarquia de subdivisão territorial de um país (cidade, estado, região, etc.).

O capítulo inicialmente descreve os conceitos utilizados na *OnLocus* para ao final descrevê-la em detalhes.

3.1 Definição da Ontologia *OnLocus*

A ontologia de lugar urbano proposta, *OnLocus*[26], é uma ontologia geográfica, classificada como uma *ontologia de extração* [Embl04] e *dependente de domínio* [Khan00]. É uma ontologia dependente de domínio porque define um conjunto de conceitos dentro de um domínio particular, o espaço geográfico urbano presente em

[26] Um sistema de referência é definido como um *Locus* ou um conjunto de *Loci* em relação ao qual as localizações espaciais são determinadas [Camp93]

páginas da Web, descrevendo feições naturais, objetos e lugares que possuem significado para uma comunidade urbana, incluindo os relacionamentos entre eles. É uma ontologia de extração porque utiliza a ontologia como suporte semântico para a extração e estruturação de dados semi-estruturados encontrados em páginas da Web. Quando aplicada a páginas da Web é capaz de reconhecer e interpretar termos referentes a lugares associando-os aos conceitos definidos na ontologia.

A *OnLocus* consiste de um conjunto *C* de conceitos (*Lugar*, *Divisão Territorial*, etc.), um conjunto *R* de relacionamentos espaciais e convencionais (topológico, "todo-parte", de localização e genérico) e um conjunto *A* de axiomas (derivados das cardinalidades associadas aos relacionamentos) usados para conceituar o domínio de interesse *D*. *D* define os locais urbanos e intra-urbanos associados a páginas da Web que fazem referência a atividades e serviços.

Uma ontologia de extração descreve regras para a identificação dos elementos da ontologia presentes nas páginas da Web como por exemplo, expressões regulares e palavras-chave indicadoras de contexto. As expressões regulares definidas na *OnLocus* para o reconhecimento de endereços estão detalhadas no Capítulo 4, nas seções 4.2 e 4.3. Já as expressões espaciais em linguagem natural que funcionam como palavras-chave na identificação de expressões de posicionamento estão definidas na seção 3.5.

A *OnLocus* pode ser representada por um grafo onde os nós representam conceitos e as arestas relacionamentos. A Figura 3.1 mostra um subgrafo desta ontologia.

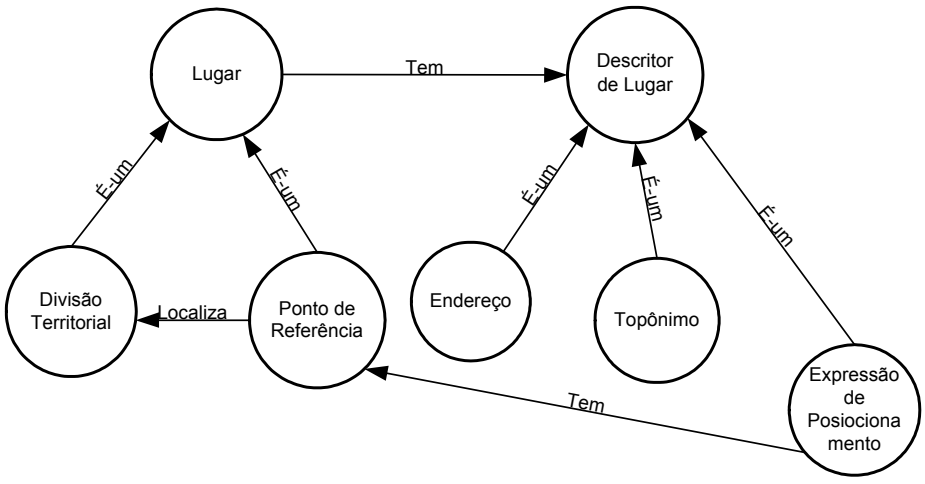


Figura 3.1 – *OnLocus* representada como um grafo

Segundo a classificação de Spaccapietra et al. [SCPV04], a ontologia *OnLocus* engloba dois tipos de ontologia: uma ontologia taxonômica e uma descritiva. A ontologia taxonômica é aqui chamada "ontologia taxonômica espacial" porque explora a estrutura hierárquica do espaço urbano, onde regiões são subdivididas em outras regiões, representada na *OnLocus* pelo conceito *Divisão Territorial* e suas especializações. A hierarquia do território está refletida nos relacionamentos mereológicos ("todo-parte") entre os conceitos que a compõem (*Cidade*, *Estado*, *Região* e *País*) e pode ser representada por um grafo $H_T = (C, r, p)$, onde C é um conjunto de nós representando os conceitos, $r \in C$ é a raiz e p: $C_i \mapsto C_j$. é uma função de herança. A Figura 3.2 ilustra esse grafo.

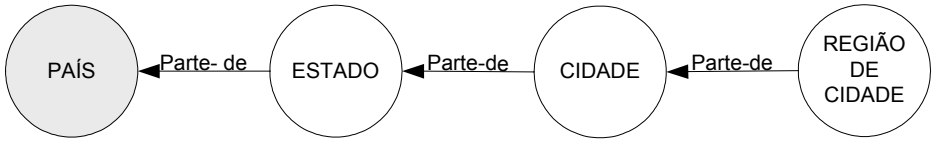


Figura 3.2 – Representação da ontologia taxonômica espacial na forma de um grafo

A ontologia descritiva conceitua formas de referenciamento a um lugar, feições naturais e pontos de referência que possuam significado de localização para uma comunidade local representado na *OnLocus* por meio dos conceitos *Ponto de Referência*, *Descritor de Lugar* e suas especializações.

3.2 Conceitos da *OnLocus*

A Figura 3.4 apresenta uma versão gráfica simplificada da *OnLocus*. Nela estão representados os principais conceitos e relacionamentos que serão descritos a seguir. A versão gráfica completa encontra-se no Anexo A junto com a sua descrição em RDF gerada pelo Protégé 2000[27] e no Anexo C, as expressões regulares definidas para o reconhecimento e extração de endereços.

Lugar é o conceito principal da *OnLocus*, sendo especializado em dois outros conceitos: *Divisão Territorial* e *Ponto de Referência*. Cada um desses conceitos é especializado em outros mais específicos. Toda instância de *Lugar* possui pelo menos uma forma de ser referenciada, representada na *OnLocus* pelo conceito *Descritor de*

[27] `http://protege.stanford.edu`

Lugar. O conceito *Descritor de Lugar* é especializado em três outros conceitos: *Endereço*, *Topônimo* e *Expressão de Posicionamento*.

A representação gráfica adotada para a ontologia *OnLocus* é baseada em algumas primitivas do diagrama de classes da *Unified Modeling Language*[28] (UML). A notação adotada, e que será usada ao longo deste capítulo, está descrita na Figura 3.3. Conceitos são representados por retângulos e os relacionamentos por linhas. Para aumentar a capacidade de representação semântica, algumas primitivas foram introduzidas para distinguir diferentes tipos de relacionamento. A notação de cardinalidade adotada na representação gráfica é a mesma empregada na UML. A cardinalidade caracteriza os relacionamentos.

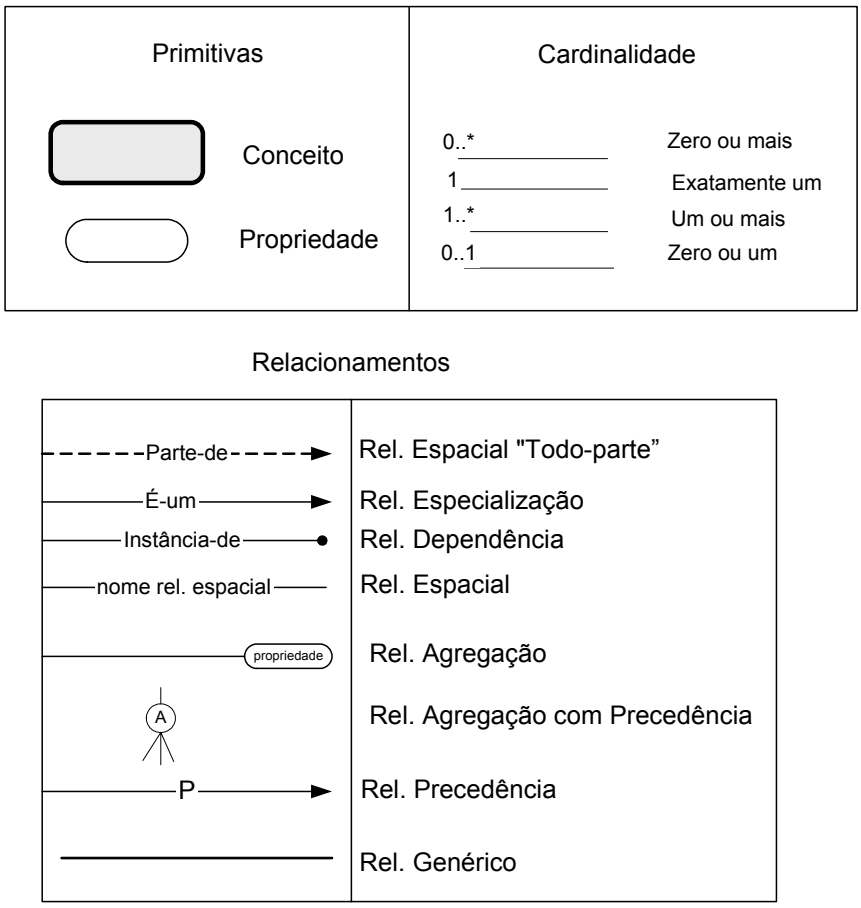


Figura 3.3- Legenda da representação gráfica

[28] http://www.uml.org

3.2.1 Lugar

O conceito *Lugar* representa descrições do espaço relacionadas a uma localização que possuem identidade própria e podem ser referenciadas de diferentes formas. Um lugar, portanto, representa mais do que uma geometria ou uma topologia: inclui um aspecto cognitivo e reflete como as pessoas percebem e usam a informação geográfica. O conceito *Lugar* é especializado em dois outros conceitos, *Divisão Territorial* e *Ponto de Referência*, que herdam a propriedade *Representação Espacial* e um relacionamento com o conceito *Descritor de Lugar* (Figura 3.4). Essa especialização não é, portanto, total, pois, além de ser uma divisão territorial ou um ponto de referência de localização, um lugar pode ser qualquer espaço ou objeto identificado por um dos descritores de lugar definidos. Como exemplo, temos uma residência que não é considerada um ponto de referência.

Cada instância de *Lugar* está associada à sua descrição geométrica. Esta descrição geométrica usa conceitos padrão em sistemas de espacialização: pontos, linhas e polígonos, estes últimos podendo ser englobados nos seus retângulos mínimos envolventes (RME).

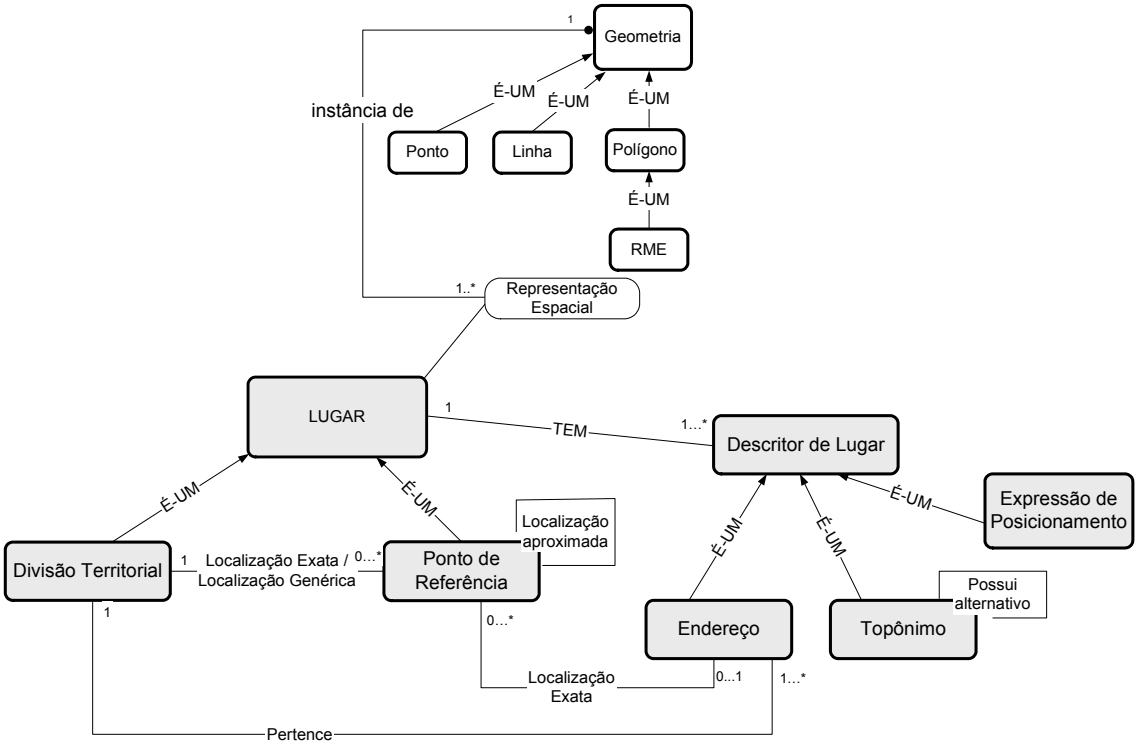


Figura 3.4- *OnLocus* - Representação gráfica simplificada

3.2.2 Divisão Territorial

O conceito *Divisão Territorial* representa o espaço territorial de um país com suas diferentes subdivisões geopolíticas, como estados, e político-adminstrativas, como divisões regionais. A Figura 3.5 refina o conceito *Divisão Territorial,* mostrando hierarquicamente todas as divisões de um território consideradas de interesse para este trabalho. A maior divisão corresponde ao conceito de *País*. Um país é subdividido em regiões que são subdivididas repetidamente até se chegar ao nível correspondente a regiões dentro de uma cidade (menor granularidade de área prevista na *OnLocus*). Outras subdivisões de *País* incluem Á*rea de CEP* e *Áreas de Código Telefônico*.

Usando a *OnLocus*, é possível inferir os relacionamentos espaciais topológicos de continência e adjacência utilizando apenas o conhecimento geográfico da estrutura hierárquica. Por exemplo, Belo Horizonte (*Cidade*) é adjacente a Nova Lima (*Cidade*). Ambas estão dentro de Minas Gerais (*Estado*), que está dentro da Região Sudeste (*Região*), que está dentro do Brasil (*País*).

3.2.3 Ponto de Referência

O conceito *Ponto de Referência* é uma especialização de *Lugar*. Representa um local, um marco urbano ou um marco ambiental reconhecido pelas pessoas e que possa ser usado para orientação espacial. Na *OnLocus*, esse conceito enumera diferentes tipos de conceito usados como referência de orientação espacial, como, por exemplo, um ponto turístico, um logradouro ou um museu (Figura 3.6). Além das propriedades herdadas de *Lugar*, o conceito *Ponto de Referência* possui uma propriedade chamada *Grau de Especificidade* que informa, para cada instância, a probabilidade dela ser única dentro de uma cidade, estado ou país. Locais identificados como únicos não possuem ambigüidade e sua localização é conhecida. Exemplos destes casos são os aeroportos, autódromos e estádios de futebol que possuem uma identificação única e conhecida.

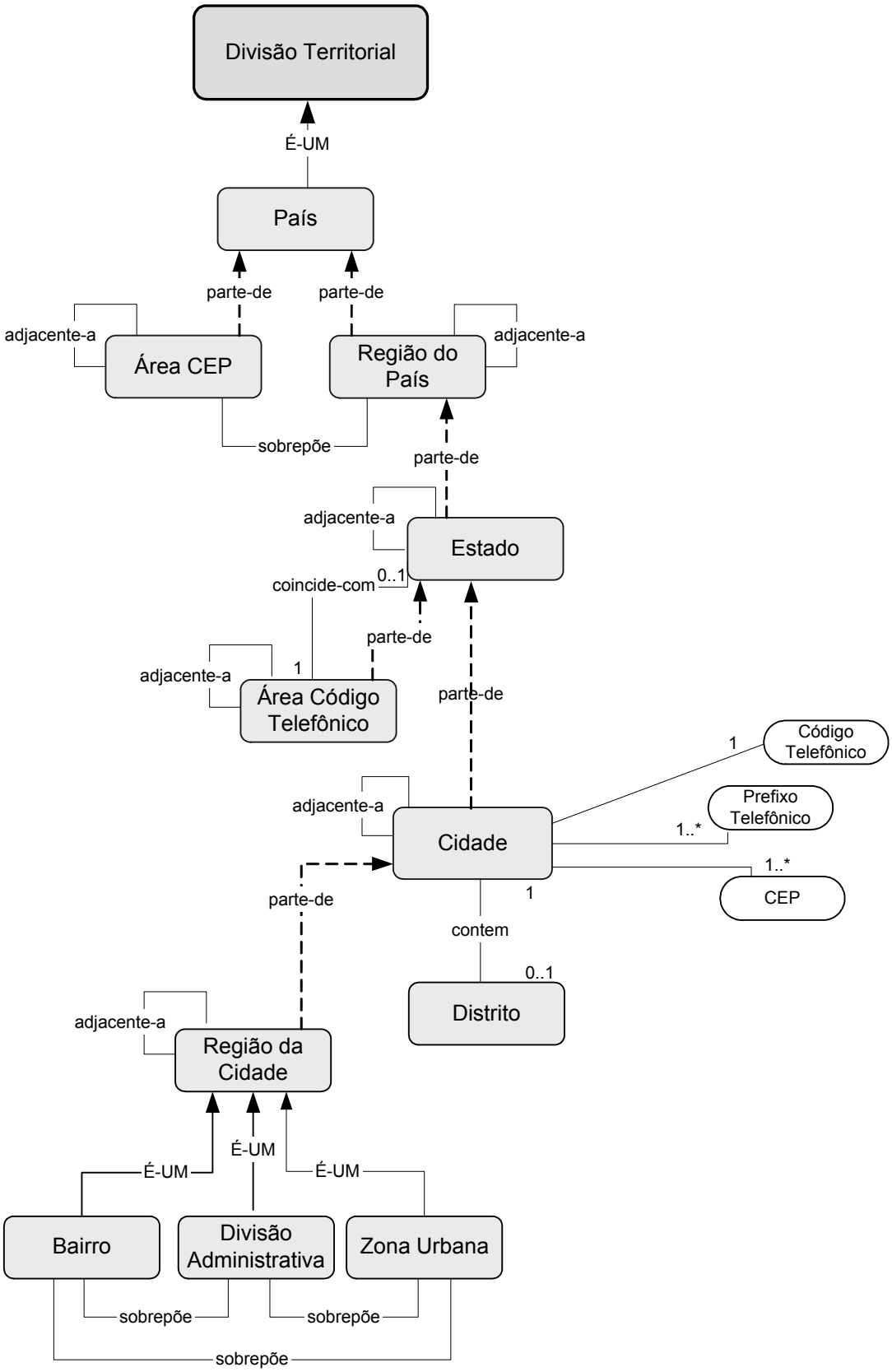


Figura 3.5 – *OnLocus* - Representação gráfica simplificada de *Divisão Territorial*.

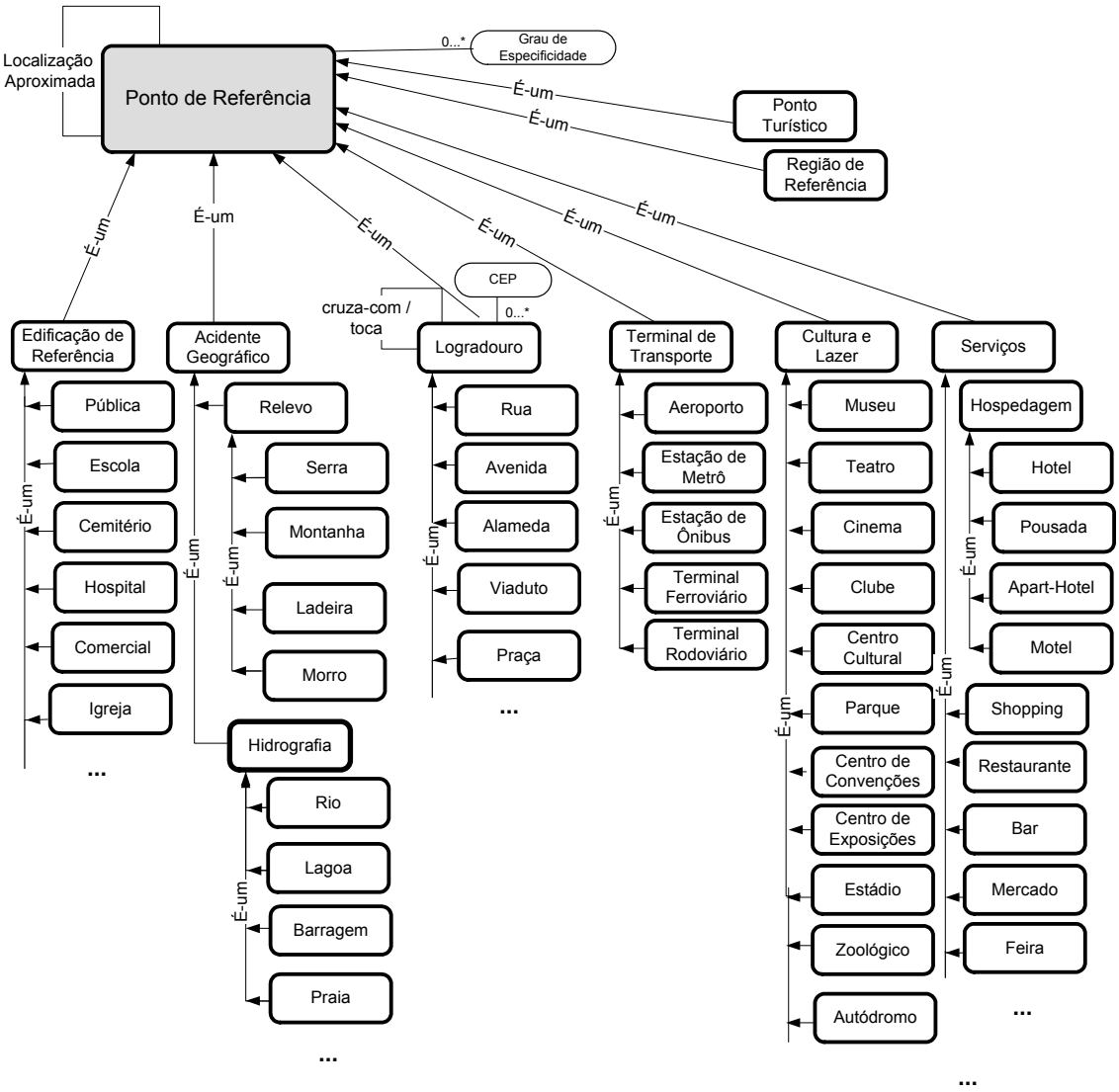


Figura 3.6 – *OnLocus* - Representação gráfica de *Ponto de Referência*

Para efeito de lugar urbano, os conceitos associados a acidentes geográficos estão suficientemente definidos na Figura 3.6. Este tipo de definição é bastante detalhado em outras ontologias, como OpenCyc ou SUMO, ou *gazetters* como TNG, e pode ser reaproveitado na *OnLocus* quando cabível. Conceitos associados a ponto de referência também estão presentes nas ontologias SUMO e SWEET. Apesar de alguns conceitos presentes nessas ontologias serem os mesmos presentes na *OnLocus*, sua posição hierárquica e seus relacionamentos apresentam diferenças principalmente devido à característica de ontologia de extração da *OnLocus*.

3.2.4 Descritor de Lugar

O conceito *Descritor de Lugar* define as formas de se referenciar um lugar. Um lugar é identificado por meio de seus descritores podendo ser um endereço, um topônimo ou uma expressão de posicionamento. Um lugar pode ser referenciado por meio de um ou mais descritores, como, por exemplo, um ponto turístico pode ser reconhecido pelo seu nome, pelo seu endereço ou por meio de uma expressão de posicionamento. A Figura 3.7 mostra a representação gráfica do conceito *Descritor de Lugar* e suas especializações.

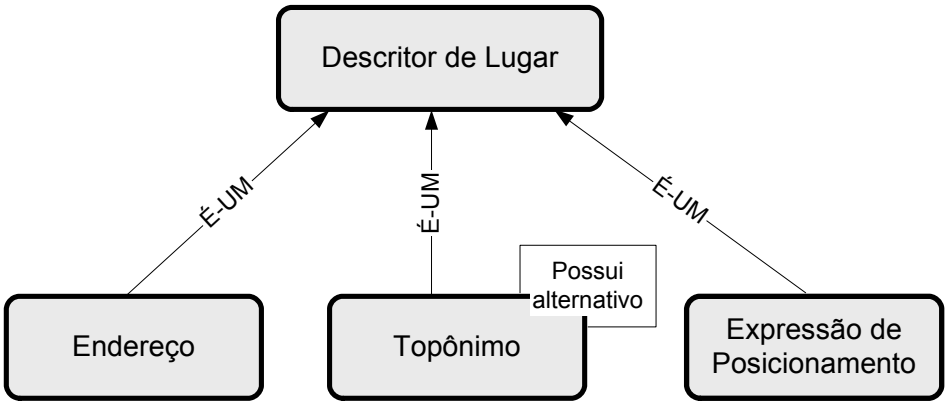


Figura 3.7 – *OnLocus* - Representação gráfica de *Descritor de Lugar*

3.2.5 Endereço

Um endereço é uma forma de expressar a localização de um lugar de interesse, sendo universalmente usado para a entrega de correspondência. Assim, na *OnLocus* o conceito *Endereço* é composto de vários elementos, que variam de acordo com o propósito pelo qual o endereço é usado e o local ao qual o endereço está associado.

A *OnLocus* considera que um endereço postal é composto de três partes: a primeira chamada de *Endereço Básico*, a segunda de *Complemento* e a terceira de *Localização*. *Endereço Básico* é composto de tipo de logradouro, nome de logradouro e número do imóvel. *Complemento* é opcional e inclui dados complementares como número do apartamento, da sala e o nome do bairro. *Localização* inclui os componentes que identificam o lugar ao qual o endereço está associado, chamados de *identificadores de localização*. A Figura 3.8 mostra a composição do conceito *Endereço* e a ordem de precedência que alguns dos elementos têm em relação a outros. O relacionamento de precedência é discutido na Seção 3.3 que trata dos relacionamentos da *OnLocus*.

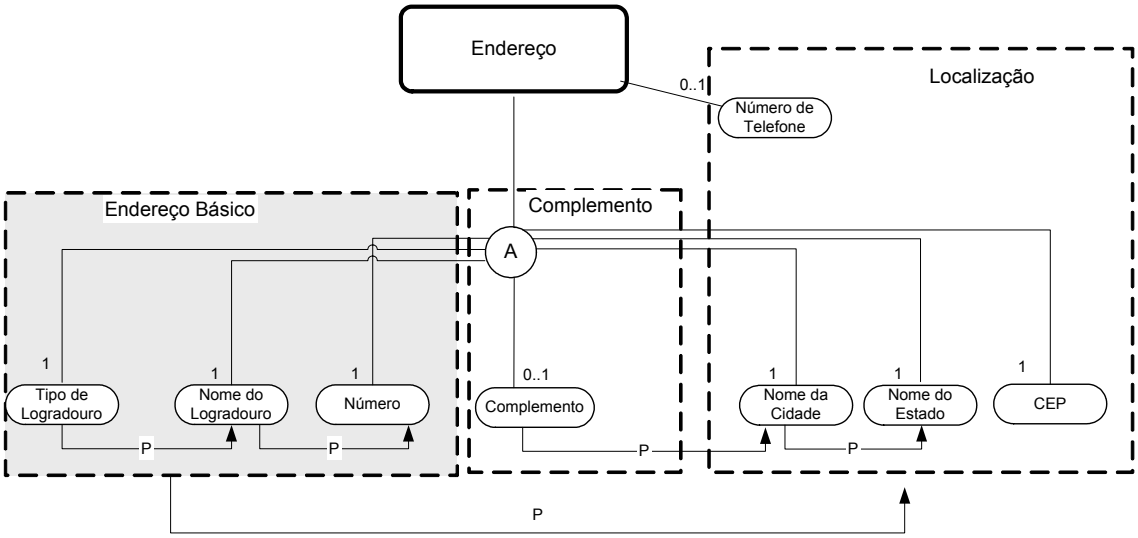


Figura 3.8 – *OnLocus* - Representação gráfica de *Endereço*

Como pode ser observado na Figura 3.8, além dos componentes de endereçamento, a *OnLocus* considera o número de telefone como propriedade de um endereço. Ele é um indicador indireto de localização e a sua presença junto a endereços em páginas da Web é tão freqüente quanto a do CEP, conforme mostrado pelos resultados dos experimentos apresentados no Capítulo 5.

A ontologia *OnLocus* considera que o endereço pode ser *completo*, *incompleto* ou *parcial*. Um *endereço completo* é aquele que possui todos os componentes utilizados em um endereçamento como: identificação do logradouro, número do imóvel, complemento do endereço, cidade, estado e CEP. Um *endereço incompleto* apresenta um conjunto mínimo de elementos utilizados em um endereçamento, suficiente para a sua localização, ou seja, o *Endereço Básico*, com ou sem *Complemento*, e algum identificador de localização (*Nome da Cidade*, *Nome do Estado, CEP* ou *Número de Telefone*). Um *endereço parcial* apresenta somente os identificadores de localização.

A ordem de aparecimento dos componentes de um endereço em páginas da Web pode variar tanto no caso de endereço completo quanto incompleto. Para serem identificados, a ordem de precedência deve ser respeitada. Assim, o *Endereço Básico* sempre deve vir em primeiro lugar, sendo necessária a existência de pelo menos um identificador de localização. O Capítulo 4 detalha as variações possíveis para a ordem de aparecimento dos componentes em um endereço, essencial para o reconhecimento de um endereço em páginas da Web.

3.2.6 Topônimo

O conceito *Topônimo* representa nomes próprios de lugares. Um lugar pode ser conhecido por vários nomes estando, portanto, sujeito ao problema de ambigüidade. Da mesma forma, um mesmo nome pode designar vários lugares distintos. Os vários nomes de um mesmo lugar estão representados na Figura 3.7 pelo relacionamento recursivo e um lugar com vários nomes pelo relacionamento entre os conceitos *Topônimo* e *Divisão Territorial ou Topônimo e Ponto de Referência.*

3.2.7 Expressão de Posicionamento

Conforme discutido no Capitulo 2, uma expressão de posicionamento é uma construção semântica em linguagem natural, muito utilizada pelas pessoas, para expressar a localização de algum lugar em função de um outro mais conhecido. Uma expressão de posicionamento é definida como um par <relacionamento espacial, ponto de referência>.

Na *OnLocus*, o conceito *Expressão de Posicionamento* é composto por uma *Expressão* que denota um relacionamento espacial seguida de um *Topônimo*. As expressões que denotam relacionamentos espaciais, em linguagem natural, consideradas pela *OnLocus* estão definidas nas Tabelas 3.2 e 3.3 (Seção 3.4). A Figura 3.9 apresenta a representação gráfica deste conceito mostrando o relacionamento de precedência entre as propriedades.

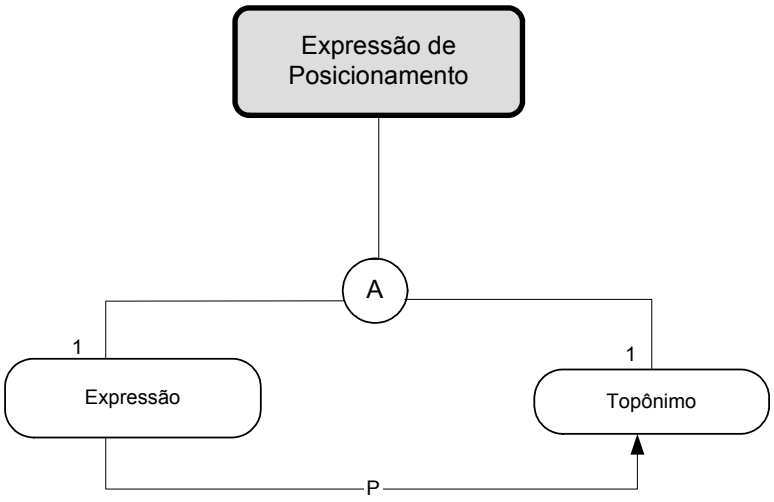


Figura 3.9 – *OnLocus* - Representação gráfica de *Expressão de Posicionamento*

3.3 Relacionamentos da *OnLocus*

A *OnLocus* considera relacionamentos binários pré-definidos, tanto convencionais quanto espaciais. Os relacionamentos convencionais ocorrem entre conceitos e também entre propriedades, sendo dos seguintes tipos: *especialização* ("é-um"), *agregação* ("parte-de") e *genérico*. Os relacionamentos espaciais ocorrem somente entre conceitos e pertencem a três tipos básicos: topológico, "todo-parte" e de localização. Os diferentes tipos de relacionamento considerados pela *OnLocus* estão descritos a seguir.

3.3.1 Relacionamentos Convencionais

3.3.1.1 Especialização[29]

Especialização é um relacionamento usado para se criar uma hierarquia de conceitos onde conceitos mais genéricos são especializados em conceitos mais específicos [SmSm77, TsLo82]. O relacionamento "é-um" ("*is-a*") é usado para expressar uma especialização. Conceitos especializados herdam todas as propriedades do conceito mais genérico, adicionando pelo menos uma propriedade que os distingue das demais especializações. Como exemplo, temos *Divisão Territorial* "é-um" *Lugar*.

3.3.1.2 Agregação[30]

Agregação é um relacionamento usado para se descrever um determinado conceito, ocorrendo entre o conceito e suas propriedades. É expresso por "parte-de", significando que o conceito é definido pelas propriedades que constituem as suas partes. Por exemplo, um *Endereço* é constituído das partes *Tipo de Logradouro, Nome do Logradouro, Número do imóvel, Complemento, Nome da Cidade, Nome do Estado, CEP e Telefone*.

Em um relacionamento de agregação pode ser aplicada uma *restrição de precedência* entre as propriedades, por meio de um relacionamento de precedência. Esta restrição define que, as propriedades de um conceito devem obedecer a uma determinada ordem estrutural para que definam o conceito. Utilizando como exemplo o

[29]Formalmente, especialização é uma abstração (primitiva) de modelagem usada em aplicações de inteligência artificial e banco de dados para a representação de hierarquia de classes, que definem relacionamentos do tipo "é-um".
[30] De forma análoga à especialização, a agregação é uma abstração (primitiva) que, ao ser aplicada a uma classe, dá origem a relacionamentos "parte-de" entre a classe e seus atributos e entre classes.

conceito *Endereço*, observamos que se algumas das propriedades definidas estiverem em uma ordem completamente aleatória, como *Nome da Cidade, Nome do Estado, Tipo de Logradouro, Número do imóvel e Nome do Logradouro*, não podemos dizer que essas propriedades caracterizam um endereço postal.

3.3.1.3 Precedência

O relacionamento de precedência é uma restrição estrutural entre propriedades dentro de uma agregação. Ele define uma seqüência padrão entre as propriedades, de forma a agrupá-las segundo uma lei de formação, explicitando uma determinada ordem Essa seqüência faz com que um conceito seja reconhecido pelas suas propriedades. Por exemplo, um endereço é considerado um *Endereço Básico* quando o *Tipo de Logradouro* vier precedido de um *Nome do Logradouro* seguido de um *Número de imóvel.*

3.3.1.4 Dependência

O relacionamento de dependência é expresso por "instância-de" e acontece quando o valor de uma propriedade de um determinado conceito é instância de outro. Por exemplo, a propriedade *Representação Espacial* do conceito *Lugar* tem como valor uma instância do conceito *Geometria*.

3.3.1.5 Genérico

Um relacionamento genérico representa uma associação entre instâncias de conceitos diferentes. É expresso por um verbo definido conforme a semântica do relacionamento. Por exemplo, o conceito *Lugar* possui o relacionamento *Tem* com o conceito *Descritor de Lugar*.

3.3.2 Relacionamentos Espaciais

Como a *OnLocus* foi proposta com o objetivo de identificar e recuperar referências a lugares presentes em páginas da Web, somente um subconjunto dos relacionamentos espaciais discutidos na literatura foi considerado. Dentre eles, destacamos os relacionamentos que denotam proximidade e continência, por serem normalmente usados pelas pessoas para descrever alguma localização e os relacionamentos "todo-parte", usados para representar a divisão hierárquica do espaço geográfico.

Os três tipos básicos de relacionamento espacial são: (1) relacionamentos *Topológicos*, (2) relacionamentos "*Todo-parte"* e (3) relacionamentos de *Localização*. Esses relacionamentos estão descritos a seguir.

3.3.2.1 Relacionamentos Topológicos

Clementine et al. [ClFO93] apresentam cinco relacionamentos topológicos distintos, que ocorrem entre objetos geográficos vetoriais dos tipos ponto, linha e polígono, e demonstram que esses relacionamentos formam um conjunto mínimo, a partir do qual todos os demais relacionamentos topológicos podem ser especificados. A definição formal desses relacionamentos topológicos usa uma notação na qual o objeto geográfico é representado por letras maiúsculas em itálico (ex., *A*, *B*), sua fronteira é representada por ∂A e seu interior por A^o (note que $A^o = A - \partial A$). A fronteira de um objeto do tipo geométrico Ponto é considerada sempre vazia e a fronteira de um objeto do tipo geométrico Linha é a parte compreendida entre os dois pontos inicial e final. A função *dim* é usada para retornar a dimensão do objeto, retornando 0 se o objeto é um ponto, 1 se o objeto é uma linha ou 2 se o objeto é um polígono (região). Os cinco relacionamentos topológicos são *toca* (*touch*), *está-contido-em* (*in*), *sobrepõe* (*overlap*), *cruza-com* (*cross*) e *disjunto* (*disjoint*). Suas definições são apresentadas a seguir:

Se *A* e *B* são objetos geográficos (regiões, pontos ou linhas) então [ClFO93]:

$A \text{ toca } B \Leftrightarrow (A^o \cap B^o = \varnothing) \wedge (A \cap B \neq \varnothing)$

$A \text{ está-contido-em } B \Leftrightarrow (A \cap B = A) \wedge (A^o \cap B^o \neq \varnothing)$

A sobrepõe B

$\Leftrightarrow \dim(A^o) = \dim(B^o) = \dim(A^o \cap B^o) \wedge (A \cap B \neq A) \wedge (A \cap B \neq B)$

A cruza-com B

$\Leftrightarrow \dim(A^o \cap B^o) = ((\max(\dim(A^o), \dim(B^o)) - 1) \wedge$
$(A \cap B \neq A) \wedge (A \cap B \neq B)$

$A \text{ disjunto } B \Leftrightarrow (A \cap B = \varnothing)$

A partir dessas definições, pode-se derivar outros relacionamentos topológicos também úteis no contexto da *OnLocus* como, por exemplo:

$A \text{ adjacente-a } B \Leftrightarrow (A \text{ toca } B) \wedge \dim(A \cap B) = 1 \wedge \dim A = 2 \wedge \dim B = 2$

$A \text{ coincide-com } B \Leftrightarrow A \cap B = A = B$

A Figura 3.10 apresenta diagramas topológicos que ilustram cada um dos relacionamentos espaciais definidos, variando os tipos de objeto envolvidos. A Figura 3.11 ilustra os relacionamentos derivados *adjacente-a* e *coincide-com*.

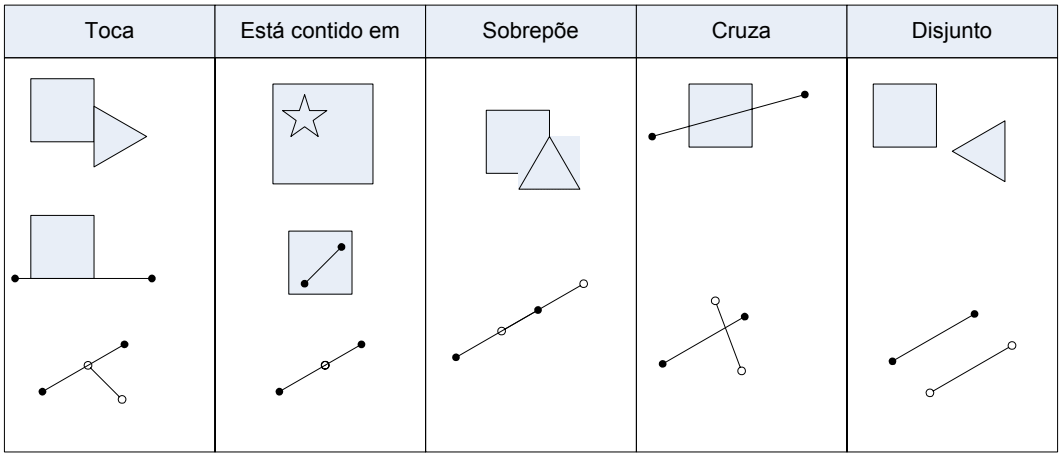


Figura 3.10– Relacionamentos topológicos definidos

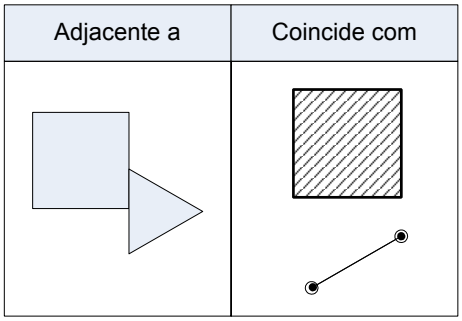


Figura 3.11- Relacionamentos topológicos derivados

É importante observar que, embora esses relacionamentos topológicos formalmente definidos tenham recebido nomes correspondentes a termos usuais em linguagem natural, seu entendimento e uso pelas pessoas não coincide necessariamente com as definições formais; por outro lado, outros termos em linguagem natural também podem ser usados para indicar os mesmos relacionamentos definidos em [ClFO93]. No caso da *OnLocus*, não é usado o relacionamento *disjunto*, o relacionamento topológico *coincide-com* é usado nos casos em que a mesma unidade espacial é usada como referência para dois ou mais conceitos e o relacionamento de *adjacência* é usado para caracterizar os casos de subdivisão planar [BoDL01], diferenciando-os dos demais casos do relacionamento *toca*.

3.3.2.2 Relacionamentos "Todo-parte"

Relacionamentos "todo-parte" são relacionamentos mereológicos entre conceitos. Um conceito representado por Cj é *parte-de* um conceito representado por Ci, se Ci tem Cj como uma parte ou Cj é *parte-de* Ci $\Rightarrow$ Cj $\subset$ Ci. O todo é mereologicamente caracterizado pela soma de suas partes. O relacionamento "*Todo-parte*" é reflexivo, assimétrico e transitivo. Por exemplo, *Município* é parte de *Estado*, que é parte de *País*. Somando-se todas as áreas correspondentes aos municípios dentro de um estado, tem-se o limite do estado. Somando-se todas as áreas correspondentes aos estados tem-se o país.

3.3.2.3 Relacionamentos de Localização

A abordagem para relacionamentos de localização aqui descrita foi inspirada no trabalho de [CaVa96] e apresenta uma versão adaptada e bem simplificada de alguns de seus conceitos. Um relacionamento de localização pode representar três tipos de localização: *exata*, *aproximada* e *genérica*. A seguir descrevemos cada um desses tipos.

Localização Exata. É um relacionamento binário comumente definido entre objetos espaciais e regiões conhecidas no espaço (georreferenciadas). Uma localização exata implica que um objeto espacial está totalmente localizado em algum local no espaço. O relacionamento de *localização exata* também pode estar descrito por meio de uma expressão de posicionamento, onde o relacionamento espacial é de continência e coincidência. A *OnLocus* usa um conjunto de expressões espaciais em linguagem natural comumente encontradas em expressões de posicionamento (ver Seção 3.4).

Localização Aproximada. É um relacionamento binário entre dois objetos espaciais, em que um objeto é localizado em função da localização de outro. Um relacionamento de *localização aproximada* acontece de duas maneiras distintas: por meio de uma localização relativa e por meio da localização aproximada de um endereço. A localização relativa acontece quando é usada uma expressão de posicionamento como, por exemplo, <hotel, perto do, aeroporto>. O hotel é localizado a partir da localização do aeroporto e da interpretação da expressão "perto do". A localização aproximada de um endereço acontece quando a posição física de um determinado endereço não é conhecida, mas é conhecida uma posição próxima. Os relacionamentos espaciais encontrados nessas expressões de posicionamento são relacionamentos espaciais que

denotam proximidade, sendo que, muitos termos são usados para representar um mesmo relacionamento espacial de proximidade. Portanto, para que um local de interesse possa ser fisicamente localizado usando uma expressão de posicionamento, é preciso se conhecer a localização física do ponto de referência utilizado e entender o significado semântico do relacionamento espacial empregado, de maneira a traduzi-lo automaticamente em uma forma de localização.

Localização Genérica. É um relacionamento binário entre objetos espaciais e regiões hierárquicas do espaço. Quando a localização exata ou aproximada não é possível, usa-se a localização genérica, significando que um objeto X está genericamente localizado em uma região Y, ou seja, alguma parte do objeto X está localizada em alguma parte da região Y. Na localização genérica não existe um comprometimento com a localização precisa, sendo apenas uma forma de identificar a qual região um objeto pertence.

3.3.3 Resumo dos Relacionamentos na OnLocus .

A Tabela 3.1 resume os relacionamentos utilizados na *OnLocus*.

Tabela 3.1 – Relacionamentos na *OnLocus*

Categoria de Relacionamentos		**Tipos de Relacionamentos**	**Forma de Expressão**
Convencional		Especialização Agregação Precedência Dependência Genérico	*é-um* *parte-de* *Precedido-por* *instância-de* Definida conforme a semântica (ex., pertence)
Espacial Quantitativo	Topológico	Adjacência Continência Sobreposição Cruza Coincidência Toca	*adjacente-a* *está-contido-em* *Sobrepõe* *cruza-com* *coincide-com* *toca*
	Mereológico	Todo-parte	*parte-de*
	Localização	Localização Exata Localização Genérica	Conforme expressões espaciais em linguagem natural de Continência e Coincidência *Localizado-em*
Espacial Qualitativo	Localização	Localização Aproximada	Conforme expressões espaciais em linguagem natural de Proximidade

Ressalta-se que há uma série de outros tipos de relacionamento espacial existentes, tanto qualitativos quanto quantitativos. Alguns dos relacionamentos espaciais da Tabela 3.1 são definidos de forma qualitativa, como, por exemplo, relacionamentos de proximidade. Experimentos descritos posteriormente nesta tese permitiram transformar tais relacionamentos em expressões que, na prática, podem ser avaliados de forma quantitativa. Em especial, vários relacionamentos da literatura não foram considerados na *OnLocus*, por não serem usuais no contexto de lugar urbano brasileiro como, por exemplo, relacionamentos de direção.

3.4 Mapeamento das Expressões Espaciais

As expressões em linguagem natural propostas para exprimir relações espaciais na *OnLocus* estão restritas ao domínio intra-urbano. Essa limitação do domínio visa diminuir ou eliminar os conflitos terminológicos dependentes de escala ao se tratar tais expressões nas páginas da Web. Conforme já mencionado, em um raciocínio espacial qualitativo existe uma tolerância de imprecisão devido à falta de dados precisos e quantitativos. Observa-se que em alguns países, como o Brasil, não é muito comum o uso de direções nas expressões de posicionamento, não sendo, por isso, consideradas nas expressões espaciais em linguagem natural definidas a seguir.

3.4.1 Mapeamento das Expressões Espaciais de Continência e Coincidência

Em linguagem natural existem várias maneiras de se denotar que um objeto espacial possui a sua localização contida ou coincidente com outro lugar. Muitas vezes, termos que denotam continência, para efeito de geocodificação, são interpretados como uma localização coincidente, como na expressão <no 3º. andar do, BH shopping>.

A Tabela 3.2 apresenta as variações semânticas mais freqüentes encontradas na Web brasileira [DeBL05, Delb05]. Dentre elas, as mais usadas são “no coração de” e “dentro de”. As expressões “no”, “na” e “em” não foram consideradas em [Delb05], mas são consideradas pela *OnLocus* por também serem freqüentes em páginas da Web brasileira. Estas expressões espaciais de continência e coincidência são comuns nas expressões de posicionamento que levam a um relacionamento de *localização exata*.

Tabela 3.2 – Mapeamento das expressões espaciais de continência e coincidência

Expressões Espaciais em Linguagem Natural	**Relacionamento Topológico Correspondente**
dentro de no coração de no n-ésimo piso de no n-ésimo andar de no n-ésimo nível de na praça de alimentação de em cima de embaixo de no / na em	*está-contido-em*, *coincide-com*

3.4.2 Mapeamento das Expressões Espaciais de Proximidade

A partir de experimentos na Web brasileira, Delboni [Delb05] atribuiu um valor métrico aproximado para cada expressão de proximidade. O que se percebeu é que, em relação ao que é considerado "próximo", existem expressões que denotam uma noção de proximidade maior que outras. No entanto, independente do valor métrico atribuído a cada expressão, existe uma equivalência entre as expressões, mostrando que todos estes termos indicam proximidade.

Esta conclusão é explorada pela *OnLocus* que adota um conjunto de expressões espaciais, em linguagem natural, que denotam proximidade e que são freqüentemente encontradas em páginas da Web brasileira [Delb05] (Tabela 3.3). Esse conjunto de expressões espaciais auxilia o reconhecimento e a interpretação do significado semântico das expressões de posicionamento encontradas em páginas da Web. Assim, baseado no trabalho de Delboni [Delb05] e usando raciocínio espacial, essas expressões foram agrupadas dentro de grupos semântica e geograficamente equivalentes. Para cada grupo foi atribuída uma região de proximidade "*default*" válida para o domínio intra-urbano. Essas regiões foram definidas com base nos valores quantitativos atribuídos aos relacionamentos métricos qualitativos obtidos através de experimentos e representam uma interpretação quantitativa aproximada. A quantificação é obtida a partir da definição de uma área em volta do ponto de referência, cujo raio corresponde ao raio do RME somado ao valor "*default*" (Tabela 3.3).

A *OnLocus* adota essa região de proximidade como uma estratégia que permite localizar geograficamente um local de interesse que não possui endereço mas tem sua localização descrita em relação a um ponto de referência de localização conhecida. Na *OnLocus* esse relacionamento é expresso como uma *localização aproximada* e, apesar de impreciso, satisfaz os requisitos de uma busca local, uma vez que, nesse tipo de busca, as consultas normalmente usam um raio para encontrar algo no entorno.

A Tabela 3.3 apresenta os vários tipos de expressão espacial de proximidade, agrupados em quatro grupos: *no entorno*, *perto, vizinhança* e *distância qualitativa*. O grupo *distância qualitativa* é um grupo especial onde, de acordo com o valor atribuído a "N", a expressão poderá ser enquadrada em um dos demais grupos de proximidade.

Tabela 3.3 – Mapeamento das expressões espaciais de proximidade

Grupo	**Expressões Espaciais de Proximidade**	**Expressões Semanticamente Equivalentes**	**Região de Proximidade Definida**
No entorno	em frente, ao lado, atrás	de fronte de, do lado	Raio do RME + 350m
Perto	perto de	pertinho de, próximo de, a poucos passos de, abaixo de, a poucos minutos de	Raio do RME + 650m
Vizinhança	nas proximidades de	na vizinhança de	Raio do RME + 1000m
Distância qualitativa	a "N" Km de a "N" metros de a "N" quadras de a aproximadamente "N" metros de a aproximadamente "N" Km de a "N" minutos de	a "N" quilômetros de a "N" m de a "N" quarteirões de, a "N" blocos de a aproximadamente "N" m de a aproximadamente "N" quilômetros de A "N" min de	Raio do RME + "N" "N" de acordo com a distância variando até 1000 m. Em minutos com N variando até 15 min (carro). Em quadras com N variando até 10 quadras

3.5 Base de Conhecimento – *Gazetteer* do *Locus*

3.5.1 Visão Geral

O *gazetteer* do *Locus*[31] [SDBD04, SDBD05, Souz05] pode ser visto como uma implementação da *OnLocus*. Esse *gazetteer* foi projetado para ser uma base de conhecimento sobre locais brasileiros, a ser utilizada no processo de reconhecimento de contexto geográfico em páginas da Web. O uso de uma base de conhecimento é necessário não só para resolver ambigüidades e validar os locais encontrados nas páginas da Web, como também para fornecer suas coordenadas geográficas e sua posição dentro da hierarquia da divisão territorial.

Os *gazetteers* existentes e utilizados como fonte de nomes de entidades geográficas do mundo inteiro, como o ADL *Gazetteer* [ADLG04], GNS [GeNS02] ou GNIS [GNIS06], não contêm informações intra-urbanas suficientes sobre o Brasil. Dessa forma, a falta de um *gazetteer* apropriado que auxiliasse o reconhecimento de lugares intra-urbanos brasileiros presentes em páginas da Web utilizando diferentes tipos de identificadores de localização, motivou a criação do *gazeetter* do *Locus*.

A criação de um *gazetteer* estruturado de acordo com a *OnLocus* resultou em um *gazetteer* mais elaborado e poderoso do que os atualmente disponíveis. Isto trouxe uma série de vantagens: (1) maior precisão semântica; (2) utilização de um número maior de relacionamentos espaciais e não só os relacionamentos de hierarquia (“parte-de”) entre regiões; (3) menor nível de granularidade espacial, podendo reconhecer entidades geográficas, como região e endereço dentro de cidade; (4) categorização de lugares intra-urbanos, como ruas, pontos de referência local, endereços, fornecendo o relacionamento entre eles; (5) uso de raciocínio espacial para derivar relacionamentos implícitos a partir de relacionamentos explícitos armazenados (por exemplo, dentro (a,b) $\wedge$ dentro de (b,c) $\Rightarrow$ dentro de (a,c)); (6) apoio na tarefa de eliminar ambigüidades (por exemplo, existem 683 instâncias “São Francisco” no *gazetteer* do *Locus*: 4 municípios, 42 bairros, 629 logradouros e 8 pontos de referência); (7) validação de nomes de lugares urbanos extraídos de páginas da Web através de identificadores diretos (como um nome de cidade) e indiretos (como um telefone); (8) integração entre a ontologia e o *gazetteer* possibilitando que modificações introduzidas

[31] `http://www.lbd.dcc.ufmg.br:8080/locus`

na ontologia possam ser refletidas no *gazetteer*; (9) acesso a dados atualizados e adequados ao contexto brasileiro.

O *Locus* estendeu a definição usual de um g*azetteer* para o que denominamos um *onto-gazetteer*, onde funções de navegação e consulta são baseadas na ontologia *OnLocus*.

3.5.2 Implementação do *Gazetteer* do *Locus*

O *gazetteer* do *Locus* implementou conceitos da *OnLocus* em três classes básicas: *Território*, *Endereço* e *Referência*, que são especializações da classe *Lugar*. No mapeamento da ontologia para o esquema conceitual do *gazetteer* optou-se por criar a classe *Endereço* como uma classe geográfica porque, no contexto de um *gazetteer*, um endereço representa uma localização geográfica dentro de uma cidade, independente de estar associado ou não a um *Ponto de Referência*. O esquema conceitual do *gazetteer* é apresentado na Figura 3.12, utilizando o modelo OMT-G para aplicações geográficas [BoDL01]. Os dados do *gazetteer* estão armazenados em um banco de dados relacional gerenciado pelo SGBD PostgreSQL[32] que inclui o módulo PostGIS[33] com funções para armazenamento espacial. Detalhes do esquema conceitual, do esquema lógico e da implementação do *gazetteer* podem ser obtidos em [Souz05].

A classe *Território* representa as divisões geopolíticas e político-administrativas brasileiras. A estrutura hierárquica das subdivisões dos territórios aparece sob a forma de agregações espaciais. Os objetos da classe *Endereço* correspondem aos endereços urbanos com localização conhecida e são armazenados individualmente (um ponto para cada endereço). O armazenamento da estrutura viária básica do município inclui os eixos de via (*centerlines)* e os cruzamentos. Locais para os quais não exista informação de endereçamento individual georreferenciado são usadas faixas de numeração associadas aos trechos de via. Um trecho de via está representado pela parte do eixo de via compreendida entre dois cruzamentos. A classe *Referência* é especializada em classes mais representativas dos lugares conhecidos e utilizados como pontos de referência pela população. Um identificador único e um nome são atribuídos a toda

[32] `http://www.postgresql.org`
[33] `http://postgis.refractions.net`

instância da classe *Lugar* (com exceção das instâncias da classe *Endereço*). Nomes alternativos ou anteriores também podem ser utilizados para identificar um local.

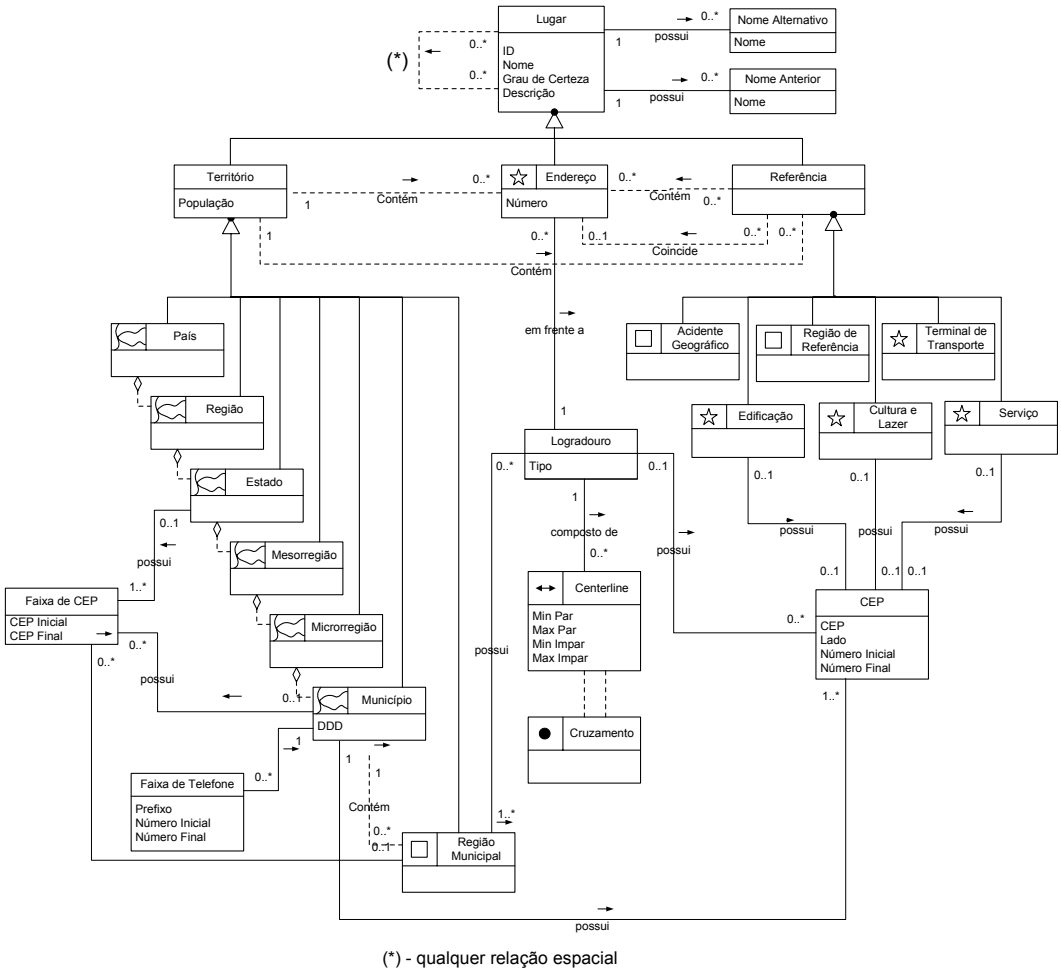


Figura 3.12 – Esquema conceitual do ga*zetteer* do *Locus*
Fonte: [Souz05]

O *Locus* prevê que vários tipos de relacionamento espacial podem ocorrer entre dois lugares quaisquer, executando, também consultas sob a forma de expressões de posicionamento [SDBD04, SDBD05, Souz05].

3.5.3 Carga do *Gazetteer* do *Locus*

Quanto maior a quantidade e qualidade dos dados em uma base de conhecimento, melhores serão as possibilidades de reconhecimento, extração e validação de contexto geográfico em páginas da Web. Por isto, para carga do *gazetteer* do *Locus* foram

necessários dois tipos principais de dados: os referentes à subdivisão do espaço geográfico do país e os relativos a pontos de referência dentro de um centro urbano.

O *gazetteer* do *Locus* foi carregado com dados das cidades brasileiras a partir de fontes de dados urbanos, regionais e nacionais, obtidos em organizações de natureza bastante diversa, como o Instituto Brasileiro de Geografia e Estatística (IBGE), a Empresa Brasileira de Correios e Telégrafos (ECT), a Agência Nacional de Telefonia (ANATEL) e instituições governamentais municipais como a Empresa de Informática e Informação do Município de Belo Horizonte (Prodabel).

Devido à origem e escala diversa dos dados obtidos, foi necessário um trabalho de compatibilização para integrá-los [Souz05]. Após a compatibilização, o *gazetteer* do *Locus* pôde inicialmente ser carregado com aproximadamente um milhão de lugares brasileiros, incluindo dados urbanos da cidade de Belo Horizonte e informações censitárias. A Tabela 3.4 mostra o número de instâncias de cada tipo.

Tabela 3.4 – Carga inicial de dados no *Locus*. Fonte: [Souz05]

Tipo	Quantidade	Fonte
País	1	IBGE
Região	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 de BH	420.942	PRODABEL
Total		1.074.777

Devido à falta de dados disponíveis e a volatilidade desse tipo de dado, outra estratégia precisou ser adotada para a carga de locais de interesse, especialmente serviços como hotéis, restaurantes, centros comerciais e lugares usados como referência de localização pela população, como monumentos, edifícios e praças. A estratégia adotada usou a própria Web como fonte de dados. Assim, para expandir o *gazetteer* do *Locus,* foram coletadas páginas de *sites* da Web contendo informações turísticas e serviços, extraindo-se os dados necessários como nome do local ou serviço, tipo de

referência, número do telefone e endereço. O experimento que mostrou a viabilidade desta abordagem está descrito no Capítulo 5 e o processo de carga correspondente está detalhado em [Souz05]. Como resultado, foram coletadas 7.741 páginas da Web brasileira, sendo o *gazetteer* do *Locus* atualizado com 29.139 novas referências. Além disso, 1.715 referências já armazenadas sofreram um ganho de informação (inserção do tipo, endereço ou telefone da referência). A Tabela 3.5 mostra o resultado da coleta e da extração de referências de páginas da Web. É importante ressaltar que a estratégia de extrair dados da própria Web atinge principalmente categorias de serviços e pontos turísticos.

Tabela 3.5 – Resultados da coleta de páginas e da extração de referências.
Fonte:[Souz05]

Endereço	**Abrangência**	**Conteúdo**	**Qte Páginas coletadas**	**Qte Referências extraídas**
cidades.terra.com.br	9 capitais e 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
Total			**7.741**	**33.315**

3.6 Conclusões

Este capítulo descreveu a ontologia de lugar *OnLocus*, classificada como geográfica, de extração e dependente de domínio. O ponto central da ontologia é o conceito de lugar incluindo lugares intra-urbanos. Além de relacionamentos convencionais, a ontologia considera relacionamentos espaciais topológicos, mereológicos e de localização. Ela foi descrita de acordo com a sintaxe do sistema *Protégé* e implementada como um *gazetteer* de um sistema de localização espacial. Após carregado, esse *gazetteer* pode ser visto como uma base de conhecimento de locais brasileiros necessária no processo de reconhecimento de lugares em páginas da Web.

O próximo capítulo discute os tipos de evidência geo-espacial que levam ao reconhecimento de um lugar em páginas da Web e as estratégias utilizadas para o seu reconhecimento e extração.

Capítulo 4

"A verdadeira viagem de descoberta
não consiste em procurar novas paisagens,
mas em ter novos olhos".
Marcel Proust

Evidências Geo-espaciais na Web

Associar um contexto geográfico às páginas da Web requer, como primeiro passo, o reconhecimento de um lugar no texto das páginas (*geoparsing*) porque essa é uma pré-condição para que a geocodificação das páginas (*geocoding*) possa ser feita. No entanto, o processo de reconhecimento de um contexto geográfico representa uma tarefa complexa, uma vez que um texto com significado geográfico pode estar em qualquer lugar da página. Além do mais, quando a evidência encontrada é um nome de lugar, o problema se torna mais complexo devido à ocorrência de ambigüidade do tipo geo/geo ou geo/não-geo [AHSS04].

Um determinado trecho do texto de uma página da Web com significado geográfico é denominado na literatura *termo geográfico* [SaKo04], *informação geo-espacial* [MAHM03], *contexto geográfico* [Mccu01], *nome geográfico* [ASS04], *referência geográfica* [RaBB03], *identificador geográfico* [Himm05] ou *referência espacial indireta* [ArSO00], e representa uma localização implícita que pode ser usada para assegurar uma correlação com uma localização geográfica. Nesta tese, essa localização implícita é denominada *evidência geo-espacial* e, conforme mencionado anteriormente, apesar de não possuírem coordenadas geográficas explícitas, as evidências geo-espaciais encontradas em páginas da Web podem ser facilmente convertidas em dados posicionais. Exemplos são nomes de regiões, códigos postais, endereços e números de telefone.

Pelo menos 20% das páginas da Web incluem uma ou mais evidências geo-espaciais reconhecíveis [Himm05]. Experimentos na Web brasileira, descritos no Capitulo 5,

mostraram que na coleção WBR05, contendo 4 milhões de páginas, 6,99% delas continham códigos postais e 12% continham números de telefone. Além disso, 3,6% dessas páginas continham uma ou mais expressões de posicionamento [Delb05] que, conforme já visto, determinam a presença de um ou mais pontos de referência.

Estas estatísticas comprovam a afirmação de Morimoto et al. [MAHM03] segundo a qual o conteúdo da Web está impregnado por informações de natureza geográfica, particularmente informação de localização como endereço postal, código postal e número de telefone. Como exemplo, os autores citam as páginas comerciais que, normalmente, contêm endereço e número de telefone de contato junto à descrição dos produtos e serviços oferecidos.

Grande parte dos trabalhos encontrados na literatura utiliza nomes de lugar como a principal evidência geo-espacial para a associação de um contexto geográfico às páginas da Web. Como observado no Capítulo 2, na tarefa de reconhecimento de nomes de lugar em páginas da Web, alguns trabalhos fazem uso de *gazetteers* [AHSS04, MKPB03, RaBB03, ZWSL05] ou ontologias [JPRS02, SMCC06] para auxiliar na resolução do problema de ambigüidade geo/geo ou geo/não-geo. O uso de outras evidências geo-espaciais, como endereço, telefone e código postal, é pouco explorado na literatura pesquisada. Os trabalhos que utilizaram esses tipos de evidência normalmente fazem uso de dados da América do Norte pela facilidade com esses dados podem ser obtidos nos EUA e Canadá.

Exemplos encontrados na literatura e citados no Capítulo 2 são relembrados aqui. A máquina de busca Geosearch [Himm05] utilizou telefones e códigos postais dos EUA e Canadá, e endereços dos EUA para geocodificar páginas da Web e, assim, possibilitar busca local. Morimoto et al. [MAHM03] utilizaram endereços e códigos postais para geocodificar páginas de serviços na região do vale do Silício, Califórnia (EUA), para uso em mineração de dados geo-espaciais. McCurley [Mccu01] usou códigos postais e números de telefone para obter a latitude e longitude de regiões dos EUA e assim geocodificar páginas da Web. Wang et al. [WXWL05] utilizaram códigos postais e telefones dos EUA para melhorar a identificação e associação de uma localização a páginas da Web e Arikawa et al. [ArSO00] utilizaram endereços como chaves espaciais para ligar páginas da Web a dados multimídia presentes na Web japonesa.

Os experimentos apresentados na literatura e os descritos no Capítulo 5 confirmam a existência de um conjunto básico de evidências geo-espaciais presentes nas páginas da Web. Cada uma dessas evidências possui sua particularidade e algumas representam indicadores mais confiáveis. De acordo com o tipo de evidência geo-espacial encontrado nas páginas da Web, somente a combinação com outros tipos definirá com mais certeza se o que foi encontrado é realmente um local ou não, evitando-se, assim, problemas de ambigüidade. Uma vez identificada uma evidência geo-espacial e solucionadas as ambigüidades semânticas que possam existir, é possível associar uma posição na superfície da terra, mesmo que de forma aproximada.

Este capítulo discute os principais tipos de evidência geo-espacial presentes na Web e as estratégias propostas para reconhecê-los, concentrando-se nos aspectos relacionados à busca local no contexto da Web brasileira.

4.1 Fontes de Evidência Geo-espacial

4.1.1 Endereços

Uma fonte óbvia de informação geográfica é o endereço já que ele é a forma natural das pessoas expressarem a localização de um lugar de interesse como local de trabalho, residência ou serviço comercial, sendo universalmente usado para a entrega de correspondência em todo o mundo. Para muitas pessoas, um endereço provê informação suficiente para a construção de uma imagem mental de localização, principalmente quando a região lhes é familiar. Outras vezes, é usado como um índice para se encontrar uma localização em um mapa [Himm05].

Páginas com endereço tendem a ter endereços confiáveis uma vez que as organizações que colocam endereço postal em suas páginas vêem a Web como um meio importante e eficaz de transmitir informação. Nos dois anos em que o Geosearch esteve disponível na Web, os pesquisadores da Vicinity observaram que 20% das páginas continham endereços americanos bem formados, ou seja, endereços onde era possível reconhecer seus componentes e efetuar uma geocodificação [Himm05]. Já os experimentos executados na Web brasileira (Capítulo 5) mostraram que 10% das páginas possuem endereços geocodificáveis.

Apesar de o reconhecimento de um endereço postal ser um problema bem estudado, principalmente em SIG, quando se trata da Web a falta de um padrão mundial é um complicador [Mccu01]. Isso se deve ao fato de os padrões de formatação variarem consideravelmente de um país para o outro, além das variações de abreviaturas, pontuações, quebra de linhas e outras características encontradas, que tornam o desenvolvimento de um *parser* (analisador sintático) para reconhecimento de endereços uma tarefa não trivial [Mccu01]. Por outro lado, dentro de um mesmo país, o reconhecimento de um endereço postal em textos em linguagem natural pode ser bem estabelecido e, segundo Himmelstein [Himm05], os endereços são computacionalmente fáceis de serem detectados.

Outra particularidade dos endereços existentes em páginas da Web é que, freqüentemente, assume-se que o usuário já sabe que o endereço é de um determinado país, omitindo-se o seu nome [Mccu01]. Na Web brasileira, observamos o mesmo comportamento com os nomes de cidades. Muitos *sites* de serviços omitem o nome da cidade e do estado ao informar um endereço porque a cidade e o estado são selecionados antes de se obter a lista de serviços. Nestes casos, o uso de outras evidências, como o número de telefone com código de área, é essencial para se saber a exata localização do endereço. A Figura 4.1 mostra uma página de exemplo do Guia de Cidades do portal Terra[34]. Observe que o nome da cidade, Belo Horizonte, aparece apenas no título da página.

Figura 4.1 – Exemplo de restaurantes em Belo Horizonte - guia de cidades do *site* Terra
Fonte: http://www.terra.com.br

[34] http://www.terra.com.br

4.1.2 Código de Endereçamento Postal

O Código de Endereçamento Postal (CEP) é uma evidência naturalmente forte de localização porque seu reconhecimento possibilita a associação direta com uma determinada região de um país. Como os experimentos descritos nesta tese utilizam apenas páginas da Web brasileira, iremos detalhar apenas a estrutura do CEP usado no Brasil.

No Brasil, o objetivo principal do CEP[35] é orientar e acelerar o encaminhamento, o tratamento e a distribuição de objetos de correspondência, por meio de sua atribuição a localidades, logradouros, unidades dos Correios, serviços, órgãos públicos, empresas e edifícios. Portanto, a partir do reconhecimento de um CEP é possível associar uma localidade e, algumas vezes, o nome de um logradouro ou edificação a um endereço.

O CEP está estruturado segundo o sistema decimal, sendo composto por Região, Sub-região, Setor, Subsetor, Divisor de Subsetor e Identificadores de Distribuição. A Figura 4.2 apresenta a estrutura do CEP. O número que o representa é composto por cinco algarismos, um hífen e três algarismos chamados de sufixo. O primeiro algarismo identifica a região, o segundo a sub-região, o terceiro o setor, o quarto o subsetor e o quinto o divisor. O sufixo é destinado à identificação individual de localidades, logradouros, códigos especiais (órgãos públicos, empresas e edifícios considerados grandes usuários) e unidades dos Correios.

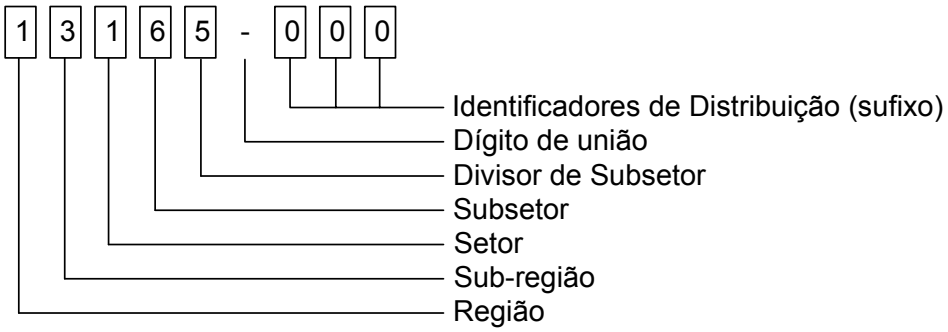


Figura 4.2 - Estrutura do CEP
Fonte: http://www.correios.com.br

Quando uma localidade possui um único CEP, não é possível identificar os logradouros individualmente e o sufixo correspondente varia de 000 a 999. As localidades que apresentam mais de um CEP (grandes municípios) possuem CEP por

35 http://www.correios.com.br

logradouro. Por exemplo, a faixa de CEP destinada a Belo Horizonte é entre 30000-000 e 31999-999. Nestes casos, o sufixo identifica um logradouro quando varia de 000 a 899, grandes usuários dos Correios quando varia de 900 a 959 ou unidades dos Correios para a faixa de 970 a 989 e 999. Existem alguns logradouros para os quais os Correios atribuem mais de um CEP, seccionando-os por faixas de numeração ou pelo lado (lado par e lado ímpar).

O Brasil é dividido em dez regiões postais para fins de codificação postal. A Figura 4.3 mostra a localização das dez regiões que correspondem ao primeiro dígito do CEP.

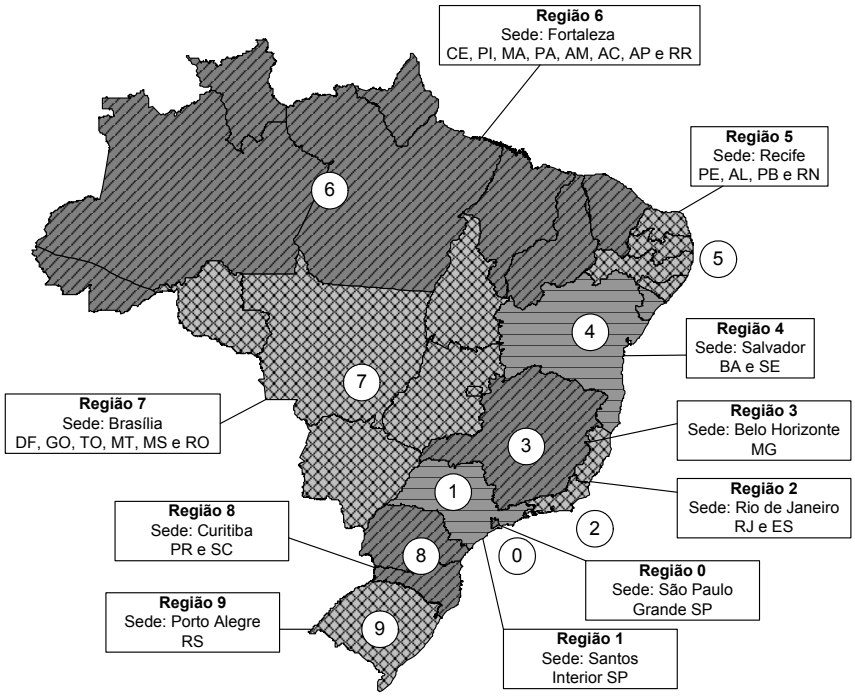


Figura 4.3 – Brasil dividido em regiões do CEP
Fonte: http://www.correios.com.br

As Figuras 4.4, 4.5, 4.6 e 4.7 exemplificam o significado de cada dígito que forma o CEP da cidade de Engenheiro Coelho (SP), apresentando sua localização geográfica dentro da codificação nacional. O primeiro algarismo representa a região postal 1, que corresponde ao interior de São Paulo. Cada região postal é divida em dez sub-regiões que são indicadas pelo segundo algarismo do CEP (Figura 4.4).

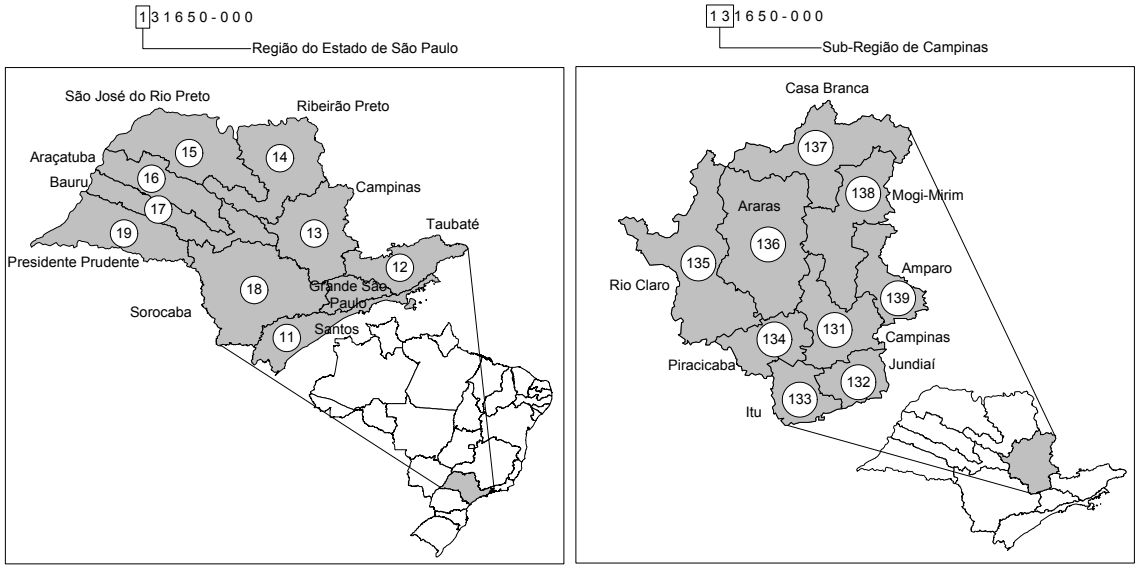


Figura 4.4 - Região postal 1 (São Paulo) e sub-regiões
Fonte: http://www.correios.com.br

No exemplo, a sub-região é a 13 cuja sede é a cidade de Campinas. Cada sub-região, por sua vez, é divida em dez setores representados pelo terceiro algarismo. No exemplo, o setor é 131 cuja sede também é a cidade de Campinas (Figura 4.5).

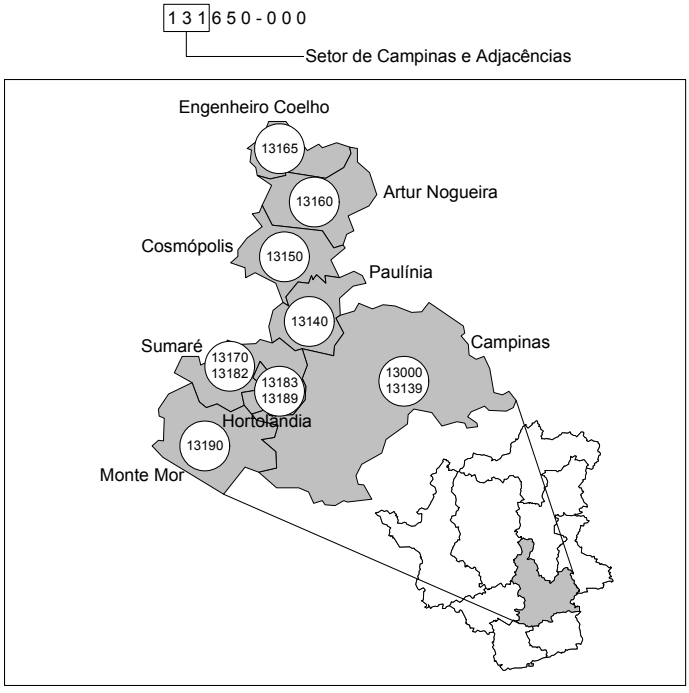


Figura 4.5 – Sub-região 13 e setores
Fonte:http://www.correios.com.br

O setor é dividido em dez subsetores representados pelo quarto algarismo. No exemplo, o subsetor é 1316 cuja sede é a cidade de Artur Nogueira (Figura 4.6).

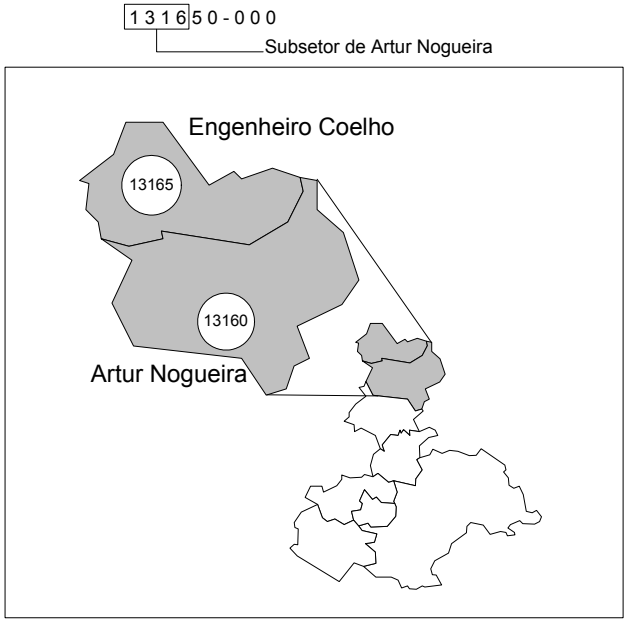


Figura 4.6 – Subsetores
Fonte:http://www.correios.com.br

Cada subsetor é divido em dez divisores representados pelo quinto algarismo. No exemplo, o divisor é 1365 cuja sede é a cidade de Engenheiro Coelho (Figura 4.7).

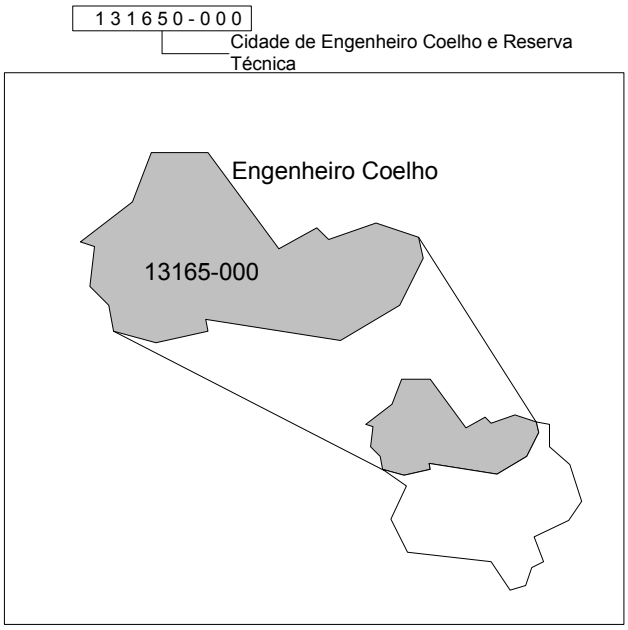


Figura 4.7 - CEP da cidade de Engenheiro Coelho (SP).
Fonte: http://www.correios.com.br

No Brasil, o CEP é um forte indicador de localização, identificando 40.835 localidades, 592.312 logradouros e 10.400 edificações, e pode ser geocodificado por diversos níveis de granularidade, como região, cidade e logradouro.

4.1.3 Números de Telefone

Os números de telefone trazem implicitamente a informação de localização [MAHM03], pois eles são organizados de acordo com princípios geográficos com o objetivo de prover um roteamento eficiente do tráfico telefônico. Por exemplo, números de telefones internacionais normalmente começam com o código do país, o que já permite associar um contexto geográfico neste nível. A União Internacional de Telecomunicações definiu o Plano de Numeração Internacional, que define o código de cada país (por exemplo, o código do Brasil é 55), assim como algumas regras básicas sobre o uso de prefixos.

À medida que se conhece a estrutura de composição dos números de telefone, mais informações são obtidas, provendo um contexto geográfico mais preciso. Descendo ao nível dos números de telefone individuais, existem bancos de dados que fornecem o endereço físico associado ao telefone. Esses bancos de dados normalmente são de propriedade das companhias telefônicas e, muitas vezes, caros de serem adquiridos devido ao seu valor comercial. Exemplos de uso do número de telefone transformado em endereço físico são comuns na área de Segurança Pública quando, nos atendimentos de emergência, um número de telefone é automaticamente traduzido em um endereço que, muitas vezes, já está mapeado em um SIG.

O reconhecimento de números de telefone em páginas da Web requer alguns cuidados para que eles não sejam confundidos com outros números como números de série, por exemplo. Padrões de codificação de números de telefone também variam dependendo se a informação fornecida é para chamada local, nacional ou internacional. Números de telefones internacionais são freqüentemente escritos seguindo um padrão que começa com "+" seguido do código do país, do código da cidade e do número do telefone. Apesar de os códigos de país e de cidade variarem de tamanho, eles são usados para fazer o roteamento das chamadas, possuindo, portanto, prefixos que identificam o país e a cidade sem dificuldades e ambigüidades. Por outro lado, como o maior tráfico de telefone é local e não internacional, muitos números são escritos de acordo com o costume local e não no padrão internacional. Também, como no endereço, existe uma grande variação no uso de parênteses, espaços e separadores na escrita dos números de telefone. A Tabela 4.1 exemplifica algumas das várias grafias para números de telefone encontradas em páginas da Web brasileira. Assim, um *parser* usado para reconhecer

números de telefone deve ser suficientemente flexível para considerar as diferentes variações de grafia. Como já mencionado, os experimentos descritos nesta tese utilizam páginas da Web brasileira. Portanto, iremos detalhar apenas a estrutura do número de telefone usado no Brasil.

Tabela 4.1 – Variações de grafia de números de telefone encontradas na Web brasileira

DDD	Número
+ 55 (31)	3377-2121
+ 55 31	33772121
(31)	3377.2121
(0xx31)	3377 2121
(0.**.31)	3377 21 21

No Brasil, o código de área (DDD) tem duas posições e divide o país em 67 áreas. A Figura 4.8 mostra a divisão do Brasil por código de DDD. O padrão para o número de telefone nacional é "0" seguido pelo código da operadora de telefonia, pelo código de área (DDD) e pelo número do telefone, que pode ter sete ou oito dígitos dependendo da região do Brasil (Tabela 4.2). Normalmente, os primeiros três ou quatro dígitos do número correspondem ao prefixo da central telefônica local à qual o assinante está conectado e os quatro últimos dígitos ao número do assinante na rede de acesso desta central. Conforme pode ser visto na Tabela 4.3, alguns estados possuem um único código de área enquanto em outros os códigos variam por cidade ou para um conjunto de cidades.

Cada cidade tem definido, além do DDD, um ou mais prefixos e uma faixa de numeração. Isto permite que o prefixo do número de telefone junto ao código de área identifique unicamente uma cidade. Por exemplo, Ouro Preto, MG, possui o DDD 31, o prefixo 3551 e os números entre 0000 e 6199. Desta forma, é possível repetir os números de telefones de forma não ambígua, em cidades diferentes.

De acordo com os experimentos descritos no Capítulo 5, o número de telefone (com código de área) é uma evidência geo-espacial com presença expressiva nas páginas da Web brasileira.

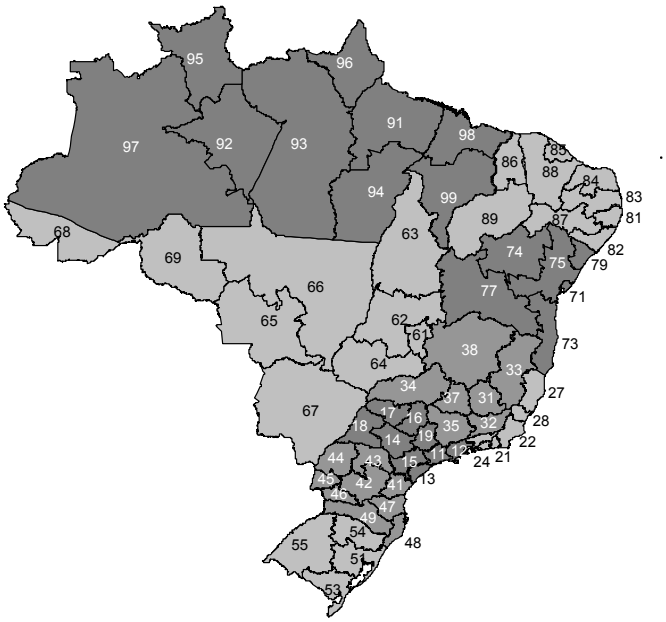


Figura 4.8 - Brasil divido por DDD
Fonte: http://www.teleco.com.br

Tabela 4.2 – Estrutura do número de telefone brasileiro

Prefixo nacional	Código da operadora	Código do DDD	Número do telefone	
			Prefixo	Número
0	NN	NN	NNN ou NNNN	NNNN

Tabela 4.3- Exemplos de códigos de DDD por estado

Estado	DDD
Acre	68
Minas Gerais	31,32,33,34,35,37,38
Goiás	61 (DF), 62, 64
Pernambuco	81,87

4.1.4 Nomes de Lugar

Um lugar aparece no conteúdo das páginas da Web como um atributo de localização de um evento ou objeto (por exemplo, o lugar onde se deu alguma notícia ou ocorreu um acontecimento) ou como caracterização do próprio conteúdo da página (por exemplo, em páginas que descrevem algum lugar específico como locais turísticos, praias ou cidades históricas).

De acordo com a *OnLocus*, um lugar pode ser reconhecido através de: (1) nomes de acidentes geográficos, como Serra do Curral, Praia de Copacabana e Lagoa da Pampulha; (2) nomes das divisões territoriais conhecidas como cidades, estados, regiões do país, regiões dos estados, bairros e regiões administrativas; e (3) nomes de logradouros e de locais de referência.

O maior problema com este tipo de evidência geo-espacial é que a maioria dos nomes de lugar encontrada na Web é ambígua [AHSS04]. Tanto ambigüidades do tipo geo/não-geo, onde o nome do lugar tem outro significado que não geográfico (por exemplo, Antônio Carlos, MG), quanto do tipo geo/geo, quando dois lugares distintos têm o mesmo nome (por exemplo, Igreja São Francisco de Assis em Ouro Preto e em Belo Horizonte), são comuns de ocorrerem no Brasil. Segundo Amitay et al. [AHSS04], 37% dos potenciais nomes geográficos mencionados nas páginas da Web possuem diversas possibilidades de significados geográficos, tendo em média dois significados por nome mencionado. Nomes de lugar podem ser imprecisos no seu significado e podem variar com o tempo e com o contexto no qual são usados. A ocorrência de variações ortográficas é muito comum como também um mesmo nome pode ser usado para identificar mais de um lugar e um lugar ser identificado por mais de um nome.

Apesar do problema de ambigüidade, os nomes de lugar são uma rica fonte de contexto geográfico que não deve ser desprezada. Como já mencionado no Capítulo 2, normalmente *gazetteers* são usados junto com um conjunto de heurísticas para se resolver as ambigüidades, fornecendo nomes e a localização de entidades geográficas.

4.1.5 Expressões de Posicionamento

As expressões de posicionamento descritas pelo par <relacionamento espacial, ponto de referência> são uma boa fonte de evidência de contexto geográfico, mas pouco explorada pela literatura [Delb05]. Através da identificação dos relacionamentos espaciais em linguagem natural, é possível o reconhecimento de locais de referência intra-urbanos. Conforme reportado em [DeBL05, Delb05], quase 82% das ocorrências de expressões de posicionamento encontradas em páginas da Web determinam a posição de um local físico em relação a um local de referência.

A vantagem desta abordagem é a eliminação da necessidade do uso de um *gazetteer* para reconhecimento de nomes, uma vez que, normalmente, em uma expressão de

posicionamento o nome de um ponto de referência vem logo após a expressão que denota um relacionamento espacial de proximidade, como nas expressões <“perto do”, “BH Shopping”> e <“1 km do”, “Aeroporto da Pampulha”>.

4.1.6 Contextos Derivados de Hyperlinks

Muitas páginas não apresentam um contexto geográfico ou ele não está claramente definido. McCurley [Mccu01] e Wang et al. [WXWL05] sugerem inferir o contexto geográfico dessas páginas através do uso de seus *hyperlinks*. Se as páginas que apontam para uma página de contexto geográfico desconhecido ou as páginas apontadas por uma página de contexto geográfico desconhecido possuírem um contexto geográfico definido, é possível inferir através da análise de seus *links* e de outras evidências a qual local a página sem contexto geográfico se refere.

Segundo McCurley, a média do número de *links* derivados de uma página é aproximadamente dez. Se 10% desses *links* apontam para páginas com contexto geográfico identificado, existe uma alta probabilidade de reconhecimento do lugar à qual a página se refere.

4.1.7 Geocodificação usando Marcadores HTML

Uma outra forma de reconhecimento de contexto geográfico é através da inserção de metadados que forneçam de forma direta o contexto geográfico. Existem propostas de uso de uma *meta tag* para codificar as informações de localização como nos padrões Dublin Core [WKLW98] e Geographic Markup Language (GML) [CDLP04]. O objetivo dessa *tag* seria permitir a especificação da latitude, longitude, altitude, região e nome do lugar ao qual a página está associada. A existência desse tipo de *tag* ofereceria vantagens desde que o autor da página entendesse o objetivo do seu uso e informasse um contexto geográfico correto. Infelizmente, McCurley [Mccu01] considera que existem inúmeros problemas com esta proposta. Primeiro, os autores precisam de ferramentas corretas para determinar as coordenadas precisas. Segundo, não existe um padrão internacional ou um registro único de nomes de lugar. McCurley, acredita que, apesar dos vários esforços existentes para especificar o formato para metadados geográficos, eles ainda não são adequados para coleções de documentos tão diversos como as páginas da Web.

4.2 Reconhecimento de Fontes de Evidência Geo-espacial

Dentre todas as evidências geo-espaciais urbanas descritas na Seção 4.1, os endereços são os mais valiosos para a busca local porque uma vez reconhecidos e analisados permitem associá-los diretamente com uma localização física urbana (geocodificação).

Como apresentado no Capítulo 2, existem várias estratégias para a extração de dados de páginas da Web. A estratégia adotada nesta tese é a baseada em ontologias. Nas ontologias de extração [Embl04], além dos objetos e relacionamentos, são definidos padrões e regras que possibilitam reconhecer e a extrair os dados de interesse. Portanto, nesta seção serão discutidas as estratégias de reconhecimento e os padrões para a extração de evidências geo-espaciais de acordo com a definição de endereço estabelecida pela ontologia *OnLocus*.

As seguintes premissas foram consideradas na definição desses padrões: (1) todo serviço na Web possui uma forma de localização associada, de modo que, se não existir um endereço associado ao serviço existe pelo menos um número de telefone que identifica o local; (2) páginas de serviços em *sites* de maior credibilidade e melhor conteúdo costumam ter informações mais bem elaboradas, incluindo um endereço bem formatado; (3) a identificação de um CEP é um indicativo da presença de um endereço.

Para evitar falsos positivos e ambigüidades foram adotadas quatro estratégias no reconhecimento de um endereço: (1) um número de telefone só é considerado se vier acompanhado do respectivo código de área (DDD); (2) um CEP que não seguir o padrão definido só será considerado se o número vier antecedido pela palavra-chave CEP; (3) um nome de cidade deve ser sempre seguido do nome ou da sigla de um estado; (4) todo endereço para ser considerado válido deverá vir acompanhado de um ou mais indicadores de localização (*Telefone*, *CidadeEstado* ou *CEP*).

A primeira estratégia evita ambigüidades entre números de telefone, a segunda evita falsos positivos entre um número qualquer e um CEP, a terceira evita ambigüidades geo/geo e geo/não-geo entre nomes de cidades e a quarta evita endereços postais de localização desconhecida. Com estas quatro estratégias, a identificação da cidade e do estado fica garantida, como veremos, a seguir, no Capítulo 5.

4.2.1 Reconhecimento de Endereços

Conforme definido na *OnLocus* (Capítulo 3), um endereço pode ser separado em três partes. A primeira parte, chamada de *Endereço Básico* (tipo e nome do logradouro, e número do imóvel) fornece o endereço propriamente dito; a segunda, chamada de *Complemento*, é opcional e corresponde ao complemento do endereço incluindo o nome do bairro; finalmente, a última, chamada de *Localização*, é subdividida em três identificadores de localização do endereço: *CEP*, *Telefone* e *CidadeEstado*. A Figura 4.9 ilustra a composição de um endereço segundo esta divisão.

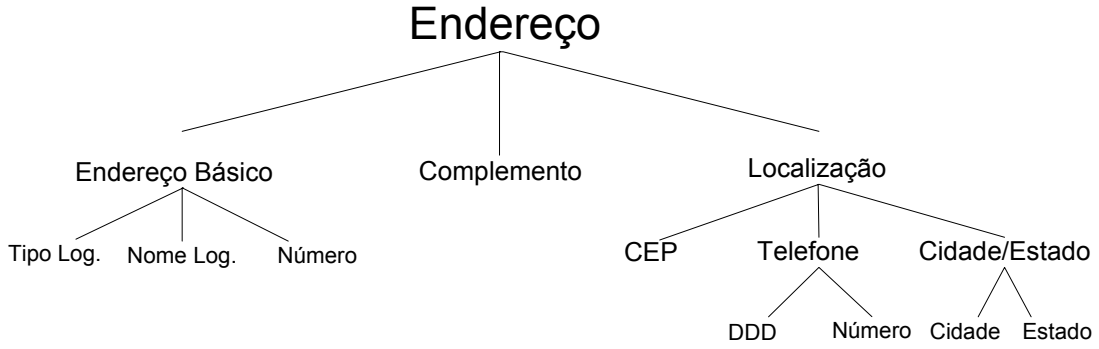


Figura 4.9 – Endereço e seus componentes

Para o reconhecimento de um endereço, apenas a primeiro parte (*Endereço Básico*) é suficiente. No entanto, para reconhecimento de sua localização pelo menos um dos três identificadores de localização (*CEP*, *Telefone* ou *CidadeEstado*) precisa existir. Quando se reconhece um endereço básico é respondida a pergunta "o que é", mas somente com a localização pode se responder "onde é".

A fim de que fossem definidos os padrões para reconhecimento de endereço, foram seguidos os seguintes passos: (1) avaliação da presença dos endereços nas páginas da Web; (2) descrição dos padrões de acordo com a avaliação feita; (3) teste da eficácia dos padrões definidos em uma coleção pequena extraída da Web brasileira; (4) seleção dos padrões mais eficazes e de maior presença na Web para teste em uma coleção de páginas representativa do domínio brasileiro (WBR05).

4.2.2 Avaliação da Presença dos Endereços

Com o objetivo de conhecer como as pessoas descrevem um endereço ou referenciam um local foi feita uma exaustiva avaliação visual em várias páginas da Web para reconhecimento das formas explícitas e implícitas de aparecimento de um lugar urbano. Como resultado, percebeu-se que para efeito de identificação e geocodificação os locais encontrados possuem duas formas de localização: absoluta e relativa.

Na localização absoluta, é identificado um endereço, um telefone, um CEP ou um nome de cidade e de estado que levarão à identificação de uma cidade em um estado e a uma geocodificação direta.

Já na localização relativa, um lugar de interesse é descrito em função da localização de outro, por meio de uma expressão de posicionamento. A Figura 4.10 exemplifica duas situações encontradas em uma página da Web brasileira onde são usadas expressões de posicionamento. Na primeira situação (1 e 2), a presença de uma localização relativa é complementar a um endereço sem identificador de localização e, na segunda (3), não existe endereço apenas a localização relativa. Em ambos os casos, conhecendo-se onde está localizado o Minascentro, é possível, a partir dele, delimitar uma área dentro da qual o hotel estaria localizado, usando como raio, o valor métrico informado.

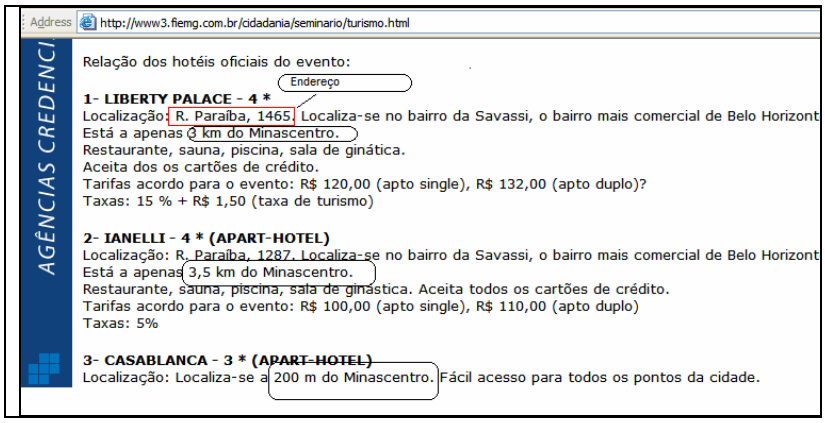


Figura 4.10 – Localização relativa em expressões de posicionamento.

É importante ressaltar que, em uma localização absoluta, a ocorrência e a ordem de aparecimento dos identificadores de localização (*CEP*, *Telefone*, *CidadeEstado*) em um endereço geralmente varia muito. Alguns endereços possuem o *Endereço Básico*

seguido pelo *Telefone*, outros seguidos pelo *CEP* ou pelo *Telefone* e *CEP*, e alguns apenas pelo identificador *CidadeEstado*. A Figura 4.11 exemplifica esses casos.

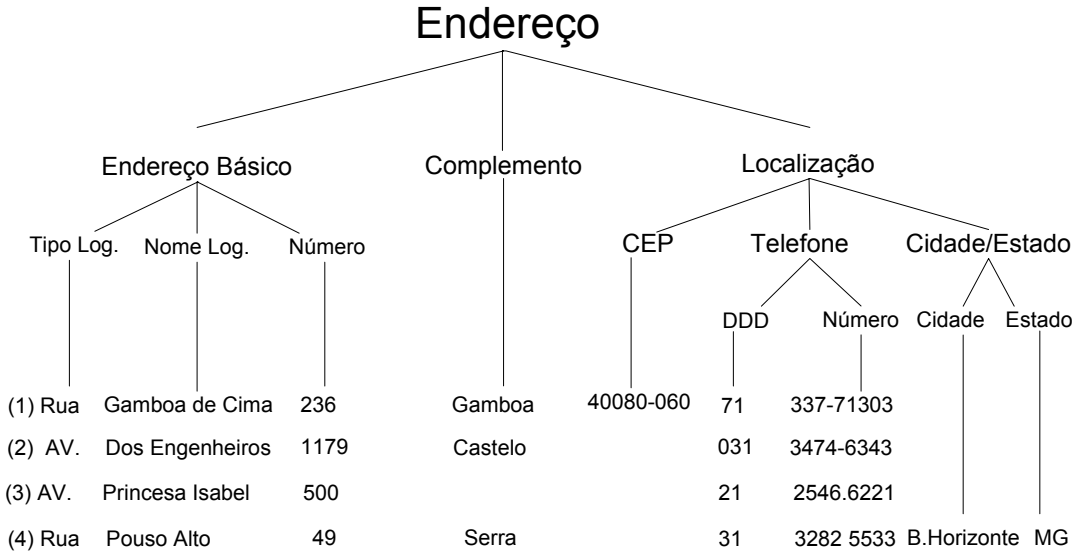


Figura 4.11 – Variações no preenchimento de um endereço

Devido às múltiplas variações na forma de se descrever um endereço, foi necessário definir padrões básicos, para cada uma das partes do endereço. Assim, através da combinação desses padrões foram criados novos padrões que foram utilizados em uma coleção de 75.413 páginas da Web brasileira com o objetivo de avaliar, entre as combinações consideradas, quais são as mais usadas e as mais confiáveis.

Esta tese trata especificamente da localização absoluta apresentando, na próxima seção, os padrões definidos para o reconhecimento de endereços em páginas da Web. Os padrões para reconhecimento de expressões de posicionamento presentes em localizações relativas, previstas na *OnLocus*, estão descritos em [Delb05] e foram desenvolvidos simultaneamente a esta tese.

4.2.3 Definição dos Padrões Básicos

Foram definidos quatro padrões básicos: (1) *Endereço Básico*, (2) *CEP*, (3) *Telefone* e (4) *CidadeEstado*. O complemento não foi definido como um padrão básico por ser opcional e variar consideravelmente. Conforme já mencionado, o texto compreendido entre o fim de um endereço básico e o início de um dos identificadores de localização foi considerado com um complemento.

Os padrões básicos são definidos a seguir utilizando a notação EBNF - *Extended Backus-Naur Form* [EBNF96]. A notação EBNF permite a definição de expressões regulares mais compactas. No Anexo B, encontra-se a notação da simbologia utilizada.

(1) Padrão *Endereço Básico*

<ENDEREÇO**>::=** [**<**IDENT>] (<TIPOLOG> <NOME> | (“BR” | “BR.”) {0-9}$^{+}$)
[<S1>] <NUM>

<IDENT>**::=** (“Endereço:” | “ENDEREÇO:”| “END:”| “End:”| “END.:”| “End.:”|
“Localização:”| “Local:”);

<TIPOLOG>**::=** (“Rua” | “RUA” | “R.” | “R” |
“Av” | “Av.” |”AV” | “AV.” | “Ave.” | “AVENIDA” | “Avenida”|
“ROD.” | “Rod.” | “ROD” | “Rod” |“Rodovia” | “RODOVIA” |
“Praça” | “PC”| “PC.”|
“Alameda” | “Al.”| “Al”|
“Travessa” | “Tv.” | “Tv”|
“Largo” | “LG.” | “Lg.” |
“Ladeira” | “Lad.” | “Lad” |
“Estrada” | “Est.” | “Est”) ;

<NOME>**::=** ({1-9} <PREP> {(A| B |...|Z) {a-z}}$^{+}$ |
{{(A| B |...|Z) {a-z}}$^{+}$ <PREP> {(A| B |...|Z) {a-z}}$^{+}$}$^{+}$ |
{(A| B |...|Z) {a-z}}$^{+}$ {1-9}|
{{(A| B |...|Z)}$^{+}$ {a-z}}$^{+}$|
{1-9}+) ;

<PREP>**::=** (“de” | “da” | “do”| “dos” | “das”| “ ”) ;

<S1>**::=** (“ , ” | “ - ” | “ ”);

<NUM>**::=** ([<N>] {(0|1|...|9)}$^{+}$ “ . ” (0|1|..|9) (0|1|..|9) (0|1|..|9) {a-z} |
{A-Z}) |
[<N>] {(0|1|...|9)}$^{+}$ ({a-z} | {A-Z}) |
“s/n” | “S/N”) ;

< N**>::=** (“n.” | “N.” |“n” | “N” | “Nº” | “nº” | “Km”| “Km.”| “KM”| “KM.”| “Número” |
“número”);

(2) Padrão *TELEFONE*

<TELEFONE>**::=** [<IDENT>] <COD-AREA> <NUMTEL>

<IDENT>**::=** (“Telefone:” | “telefone:” | “TELEFONE” | “Tel:” | “tel:”| “Tel. :”|
“tel. :”| “TEL” | “tel”| “TEL.” | “Fone.” | “Fone:”| “fone:” |
“Fone/Fax:” | “fone/fax:” | “FONE/FAX” | “TELFAX:” |
“Telefax:” | “FAX” | “Fax:”) ;

<COD-AREA>**::=** [“+55”] [“(”] (“0xx” | “0.xx.” | “0XX” | “0.XX.” | “0.**.” | “0”)
(1|2|...|9) (0|1|...|9) [“)”] ;

<NUMTEL>**::=** <PREFIXO> [<S3>] <SUFIXO>

<PREFIXO>**::=** (1|2|...|9) (0|1|...|9) (0|1|...|9) [0-9)] ;

<S3**>::=** **(**“.” | “-” | “ ”);

<SUFIXO**>::=** (0|1|...|9) (0|1|...|9) (0|1|...|9) (0|1|...|9) ;

(3) Padrão *CEP*

<CEP>**::=** ([<IDENT>] <PREFIXOCEP> <S2> <SUFIXOCEP> |
<IDENT> <PREFIXOCEP> <SUFIXOCEP>);

<IDENT>**::=** (“CEP:” | “Cep:” | “CEP” | “Cep” | “CEP.” | “Cep.”) ;

<PREFIXOCEP>**::=** (0|1|...|9) (0|1|...|9) [“.”] (0|1|...|9) (0|1|...|9) (0|1|...|9) ;

<S2>**::=** (“-” | “ . ”) ;

<SUFIXOCEP**>::=** (0|1|...|9) (0|1|...|9) (0|1|...|9) ;

(4) Padrão *CidadeEstado*

<CidadeEstado>**::=** (<SEP3> < CIDADE> <S4> <ESTADO> |
<SEP3> < CIDADE> “(” <ESTADO> “)”)

<CIDADE>**::=** ({(A|B|..|Z) {a-z}}$^{+}$ |
{{(A| B |...|Z) {a-z}}$^{+}$ <PREP> {(A|B |...|Z) {a-z}}$^{+}$}$^{+}$)

<PREP>**::=** (“de” | “da” | “do”| “dos” | “das”| “ ”) ;

<ESTADO>**::=** (<NOMEST > | <SIGLA>) ;

<NOMEST**>::=** (“Acre” | “Alagoas” | .....| “Tocantins”); (* nomes dos estados do Brasil e abreviaturas como “M. Gerais” *)

<SIGLA>**::=** (“AC” |..... | “TO”) ; (* Siglas dos estados do Brasil *)

<SEP3>**::=** “-” | “ , ” | “ ”| “ . ” | “Cidade: ” | “cidade: ” | “ <S>” ; (* <S> representa os marcadores HTML que na coleção usada foram substituídos por <S> *)

<S4>**::=** “ - ”| “ , ” | “ ” | “ / ” ;

4.3 Padrões e Regras para Extração de Endereços

Uma vez estabelecidos os padrões básicos para reconhecimento de um endereço, foram definidas regras de extração que permitem descrever como ter acesso à informação a qual estamos interessados.

De acordo com as premissas consideradas (Seção 4.2), algumas regras foram consideradas na definição dos novos padrões: (1) em um endereço completo, o padrão *Endereço Básico* ocorre sempre em primeiro lugar; (2) a variação na ordem de aparecimento dos identificadores de localização em um endereço deve ser considerada; (3) todo padrão de endereço definido deve considerar pelo menos um identificador de localização; (4) os endereços reconhecidos pelos padrões adotados são considerados evidências geo-espaciais em potencial até que sejam validados. Se o endereço extraído não contiver pelo menos um de seus identificadores de localização existente no *gazetteer* do *Locus*, não será possível validar de onde ele é, sendo, portanto, desconsiderado, já que não pode ser geocodificado. O objetivo da validação após a extração é agilizar a fase de reconhecimento e extração.

De acordo com a *OnLocus,* um endereço completo consiste do *Endereço Básico*, seguido de *Telefone*, *CEP* e *CidadeEstado*. Já um endereço incompleto consiste do *Endereço Básico* seguido de um ou dois identificadores de localização (*Telefone*, *CEP* ou *CidadeEstado*) e um *Endereço Parcial*, só possui identificadores de localização.

Com a combinação dos padrões básicos já descritos, novos padrões de endereço completo, incompleto e parcial foram gerados e, a partir deles, expressões regulares

utilizadas para o reconhecimento e a extração de endereços. A seguir são apresentadas as expressões que definem os padrões para as diversas variações consideradas para um endereço.

(1) Endereço Completo

EC.1 - Endereço Básico + CidadeEstado + CEP + Telefone

EC.1**::=** <ENDEREÇO> [<SEP1> < Compl>] < SEP3 > <CidadeEstado> <SEP2>
<CEP> <SEP2> <TELEFONE> ;

<SEP1>::= " - "| " , " | " / " | " " | "<S>" ;
<SEP2>::= "-"| ", " | " " | "<S>" | "." ;
<SEP3> **::=** "-"| ", " | " " | ". " | "Cidade: " | "cidade: " | " <S>" ;

EC.2 - ***Endereço Básico + CEP + CidadeEstado + Telefone***

EC.2**::=** <ENDEREÇO> [<SEP1> < Compl>] <SEP2> <CEP>< SEP3 >
<CidadeEstado> <SEP2> <TELEFONE> ;

<SEP1>::= " - "| " , " | " / " | " " | "<S>" ;
<SEP2>::= "-"| ", " | " " | "<S>" | "." ;
<SEP3> **::=** "-"| ", " | " " | ". " | "Cidade: " | "cidade: " | " <S>" ;

(2) Endereço Incompleto

EI.1 - ***Endereço Básico + Telefone***

EI.1**::=** <ENDEREÇO> [<SEP1> < Compl>] <SEP2> <TELEFONE>

<SEP1>::= " - "| " , " | " / " | " " | "<S>" ;
<SEP2>::= "-"| ", " | " " | "<S>" | "." ;

EI.2 - ***Endereço Básico + CidadeEstado + CEP***

EI.2**::=** <ENDEREÇO> [<SEP1> < Compl>] <SEP3> <CidadeEstado> <SEP2>
<CEP>

<SEP3> **::=** "-"| ", " | " " | ". " | "Cidade: " | "cidade: " | " <S>" ; (* <S> representa os marcadores HTML que na coleção usada foram substituídos por <S> *)

<SEP2>::= "-"| ", " | " " | "<S>" | "." ;

EI.3 - Endereço Básico + CEP + CidadeEstado

EI.3**::=** <ENDEREÇO> [<SEP1> < Compl>] <SEP2> <CEP> <SEP3>
<CidadeEstado>;

<SEP1>::= “ - ”| “ , ” | “ / ” | “ ” | “<S>” ;
<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;
<SEP3> ::= “-”| “, ” | “ ” | “. ” | “Cidade: “ | “cidade: ” | “ <S>” ;

EI.4 - ***Endereço Básico + CidadeEstado***

EI.4**::=** <ENDEREÇO> [< SEP1 > < Compl>] < SEP3 > <CidadeEstado> ;

<SEP1>::= “ - ”| “ , ” | “ / ” | “ ” | “<S>” ;
<SEP3> **::=** “-”| “, ” | “ ” | “. ” | “Cidade: “ | “cidade: ” | “ <S>” ;

EI.5 - ***Endereço Básico + CEP***

EI.5**::=** <ENDEREÇO> [<SEP1> < Compl>] <SEP2> <CEP> ;

<SEP1>::= “ - ”| “ , ” | “ / ” | “ ” | “<S>” ;
<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

EI.6 - ***Endereço Básico + CEP + Telefone***

EI.6**::=** <ENDEREÇO> [<SEP1> <Compl>] <SEP2> <CEP> <SEP2> <TELEFONE>;

<SEP1>::= “ - ”| “ , ” | “ / ” | “ ” | “<S>” ;
<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

EI.7 - ***Endereço Básico + Telefone + CEP***

EI.7**::=** <ENDEREÇO> [<SEP1> <Compl>] <SEP2> <TELEFONE> <SEP2> <CEP>;

<SEP1>::= “ - ”| “ , ” | “ / ” | “ ” | “<S>” ;
<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

EI.8 - ***Endereço Básico + CidadeEstado + Telefone***

EI.8**::=** <ENDEREÇO> [<SEP1> <Compl>] <SEP3> <CidadeEstado> <SEP2>
<TELEFONE> ;

<SEP1>::= “ - ”| “ , ” | “ / ” | “ ” | “<S>” ;

<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

<SEP3> **::=** “-”| “, ” | “ ” | “. ” | “Cidade: ” | “cidade: ” | “ <S>” ;

EI.9 - ***Endereço Básico + Telefone+ CidadeEstado***

EI.9**::=** <ENDEREÇO> [<SEP1> <Compl>] <SEP2> <TELEFONE> <SEP3>
<CidadeEstado> ;

<SEP1>::= “ - ”| “ , ” | “ / ” | “   ” | “<S>” ;

<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

<SEP3> **::=** “-”| “, ” | “ ” | “. ” | “Cidade: ” | “cidade: ” | “ <S>” ;

(3) Endereço Parcial

Existem casos onde os padrões definidos não conseguem reconhecer um *Endereço Completo* ou *Incompleto*, mas conseguem reconhecer os identificadores de localização. Como forma de atender a esses casos, também foram definidos padrões para encontrar apenas parte do endereço que corresponde a um identificador de localização, denominado nesta tese de *Endereço Parcial*. Assim, utilizando apenas *CidadeEstado*, *CEP* ou *Telefone* pelo menos o reconhecimento da cidade e do estado presentes nas páginas da Web é garantido. Nesses padrões, foi considerado cada um dos identificadores de localização, como também a combinação de dois deles que, conforme será visto no Capítulo 5, permite verificar a qualidade da informação, confirmando se os dois identificam um mesmo lugar, e avaliar qual deles é mais confiável para geocodificação.

EP.1 - CidadeEstado + Telefone

EP.1**::=**<CidadeEstado> <SEP2> <TELEFONE>

<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

EP.2 - Telefone+ CidadeEstado

EP.2**::=**<TELEFONE> <SEP3> <CidadeEstado>

<SEP3> **::=** “-”| “, ” | “ ” | “. ” | “Cidade: ” | “cidade: ” | “ <S>” ;

EP.3 - CEP + Telefone

EP.3**::**=<CEP> <SEP2> <TELEFONE>

<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

EP.4 - Telefone + CEP

EP.4**::**=<TELEFONE> <SEP2> <CEP>

<SEP2>::= “-”| “, ” | “ ” | “<S>” | “.” ;

EP.5 - Telefone

EP.5**::**=<TELEFONE>;

EP.6 - CEP

EP.6**::**=<CEP>;

EP.7 - CidadeEstado

EP.7**::**=<CidadeEstado>;

4.4 Seleção dos Padrões mais Eficazes

Para se identificar os padrões de extração mais eficazes, foi realizado um experimento preliminar no qual cada um dos dezoito padrões definidos foi testado em uma coleção de 75.413 páginas da Web brasileira para que fosse possível se ter um retrato do uso de endereços na Web. Neste experimento, procurou-se responder às seguintes perguntas: (1) Qual a forma mais comum de se encontrar um endereço descrito na Web brasileira? (2) Qual dos identificadores de localização aparece com mais freqüência? (3) Qual dos identificadores de localização é mais confiável para geocodificação? (4) Qual o percentual de aparecimento dos diversos padrões na Web? O objetivo deste experimento foi reduzir o número de padrões a serem submetidos a uma coleção maior, evitando padrões redundantes, pouco eficazes ou com pouca representatividade.

A análise dos resultados obtidos neste experimento está detalhada no Capítulo 5. De acordo com esses resultados, os dezoito padrões definidos foram inicialmente reduzidos para onze, correspondentes aos padrões que obtiveram maior sucesso na extração de endereços, e depois para os seis padrões finais. Esses seis padrões representam o conjunto que permite efetuar o reconhecimento e a extração de endereços em páginas da Web e associar um lugar ao endereço extraído. A Tabela 4.4 apresenta o primeiro conjunto de onze padrões de endereço selecionado e que representa 99,64% das ocorrências na coleção de 75.413 páginas. A Tabela 4.5 apresenta o conjunto final selecionado (6 padrões) para o reconhecimento e extração de endereços.

Observou-se, nos experimentos descritos no Capítulo 5, que a presença de um dos identificadores de localização – *CEP* ou *Telefone* – garante uma geocodificação mais eficiente por se tratar de um padrão numérico ao contrário de *CidadeEstado* que, por ser somente texto, está mais propenso a erros de reconhecimento além de erros de grafia e diferentes abreviaturas. Dessa forma, a escolha dos seis padrões privilegiaram os que incluem *CEP* e *Telefone*.

Tabela 4.4 – Padrões de endereço mais eficazes

Tipo de Padrão		**Total Extraído**	
EI.1	*Endereço Básico + Telefone*	26.350	2,9%
EI.2	*Endereço Básico + CidadeEstado+ CEP*	6.981	0,8%
EI.3	*Endereço Básico + CEP+ CidadeEstado*	21.424	2,4%
EI.4	*Endereço Básico + CidadeEstado*	19.067	2,13%
EI.5	*Endereço Básico + CEP*	43.074	4,8%
EI.6	*Endereço Básico + CEP + Telefone*	1.860	0,2%
EP.2	*Telefone + CidadeEstado*	858	0,1%
EP.3	*CEP + Telefone*	2.358	0,3%
EP.5	*Telefone*	151.641	17%
EP.6	*CEP*	99.842	11,2%
EP.7	*CidadeEstado*	516.413	57,81%
	Sub-total	889.868	99,64%
	Total de Endereços Reconhecidos	893.260	100%

Tabela 4.5 – Padrões de endereço selecionados

Tipo de Padrão	
EI.1	*Endereço Básico + Telefone*
EI.2	*Endereço Básico + CidadeEstado+ CEP*
EI.4	*Endereço Básico + CidadeEstado*
EI.5	*Endereço Básico + CEP*
EP.5	*Telefone*
EP.6	*CEP*

Para a redução dos onze padrões a seis, foi feita uma análise da geocodificação obtida com cada padrão e verificado se os padrões com maior número de componentes poderiam ser substituídos por padrões menores e mais eficazes. Assim, os padrões EI.3 e EI.6 foram excluídos por terem a parte principal (*Endereço Básico* + identificador de localização) contemplada pelo padrão EI.5. Além do mais, a geocodificação usando o *CEP* obteve um percentual maior do que usando o *Telefone* e *CidadeEstado*. A permanência desses dois padrões só se justificaria se o *Telefone* e *CidadeEstado* fossem usados como uma evidência adicional quando, por algum motivo, o *CEP* não fosse encontrado no *gazetteer* do *Locus*. O padrão EP.3 está contemplado nos padrões EP.5 e EP.6 e, portanto, foi eliminado. O padrão EP.5 contempla o padrão EP.2. Apesar de o padrão EI.4 corresponder a parte do padrão EI.2, ele foi mantido por conter o *CEP*, o que propicia maior confiabilidade na geocodificação. O padrão EP.7 foi eliminado por gerar muitos falsos positivos como, por exemplo, quando um nome próprio ou substantivo que não designa uma cidade precede o nome ou a sigla de um estado (ex., "Universidade Federal de Minas Gerais/**MG**", "Casas **Bahia**", "Festa do Divino **Espírito Santo**"). Também é comum a presença deste padrão em contextos que não constituem endereços como, por exemplo, uma página sobre a cidade de Belo Horizonte (Figura 4.12). Apesar de ter obtido o maior percentual de extração, o padrão EP.7 apresentou o pior percentual de geocodificação, já que apenas 50,76% das ocorrências foram consideradas corretas (Capítulo 5). Assim, os seis padrões selecionados foram: (1) *Endereço Básico* + *Telefone*, (2) *Endereço Básico* + *CidadeEstado* + *CEP*, (3) *Endereço Básico* + *CidadeEstado*, (4) *Endereço Básico* + *CEP*, (5) *Telefone* e (6) *CEP* (Tabela 4.5).

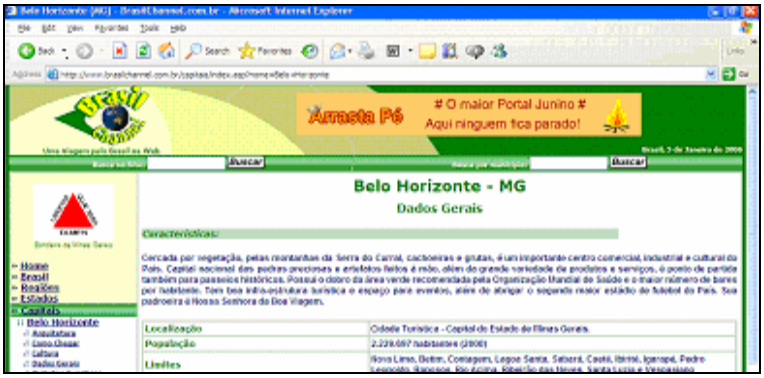


Figura 4.12 – Página sobre Belo Horizonte

4.5 Conclusões

Neste capítulo, foram discutidos os tipos de evidência geo-espacial presentes na Web e quais, dentre eles, são importantes para o contexto da *Web local*. Um dos principais objetivos para se efetuar o reconhecimento dessas evidências é, entre outras aplicações, possibilitar às máquinas de busca mecanismos para efetuar busca por local e por proximidade. A abordagem adotada para se efetuar esse reconhecimento tem como base os endereços encontrados nas páginas da Web e considera dois níveis de granularidade: municipal e local (endereço dentro de uma cidade). Assim, é garantido que pelo menos a cidade e o estado onde um serviço ou atividade está localizado sejam identificados e a cidade geocodificada. Foram apresentados os padrões mais comuns para a descrição de um endereço e, com base em experimentos, definido um conjunto mínimo de padrões que contempla as formas mais utilizadas para a descrição de um endereço na Web brasileira.

No Capítulo 5, a seguir, são apresentados os resultados experimentais que validaram nossas hipóteses e definiram esse conjunto mínimo de padrões para reconhecimento e extração de endereços.

Capítulo 5

"The best theory is inspired by practice.
The best practice is inspired by theory".
Donald E. Knuth

Resultados Experimentais

Para demonstrar a efetividade das idéias propostas nesta tese, foi realizada uma série de experimentos tendo os seguintes objetivos: (1) conhecer as formas de endereçamento usadas em páginas da Web e (2) avaliar a viabilidade de extração e geocodificação automática de endereços encontrados em páginas da Web.

Os experimentos realizados, descritos a seguir, utilizaram páginas da Web brasileira, entre elas, aquelas que compõem a coleção WBR05 com 4 milhões de páginas, dados armazenados no *gazetteer* do *Locus* e o banco de dados geográfico de Belo Horizonte contendo, além de dados sobre logradouros, bairros, endereços, pontos de referência e divisões administrativas, imagens de 40 cm de resolução da cidade.

5.1 Experimento 1 – Reconhecimento de Endereços em Páginas da Web

Este experimento preliminar teve como objetivo principal avaliar a viabilidade de se geocodificar automaticamente endereços extraídos de páginas da Web. Como forma de validar a idéia proposta, foi feita a extração de dados de localização expressos sob a forma de endereços de Belo Horizonte. A Figura 5.1 ilustra o processo envolvido no experimento e que pode ser resumido em cinco etapas básicas: (1) coleta das páginas contendo dados de interesse, (2) extração dos endereços e de dados relevantes e disponibilização no formato XML, (3) geocodificação dos endereços para a obtenção das coordenadas geográficas, (4) atualização do banco de dados geográfico usando as coordenadas obtidas, e (5) integração dos dados extraídos em um único ambiente geográfico.

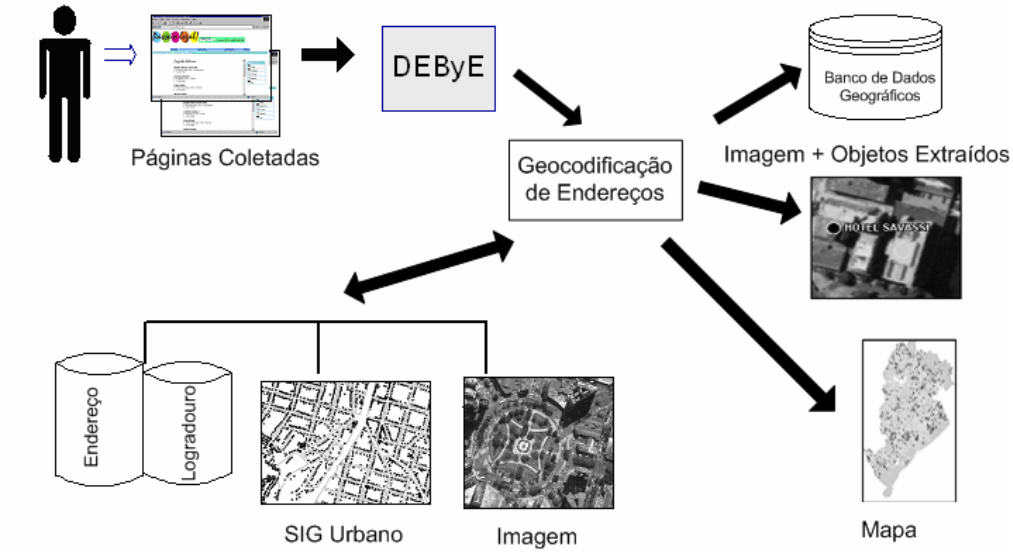


Figura 5.1 – Visão geral do processo usado no experimento

Os dados utilizados na aplicação vieram de seis fontes distintas: imagens de alta resolução de Belo Horizonte[36], dados do banco de dados geográfico urbano de Belo Horizonte e quatro *sites* da Web (Passeio Legal[37], Comida di Buteco[38], portal do Terra[39] e In BH[40]). Os *sites* selecionados provêm informação referente a serviços e atividades como hotéis, restaurantes, centros culturais, museus, bares, consulados, agências de publicidade, casas de espetáculo, floriculturas, joalherias, lojas de moda feminina, cinemas, bibliotecas e hospitais. Um subconjunto de páginas disponíveis em cada *site* foi coletado totalizando 65 páginas e 540 objetos[41].

Para a extração de dados, nomes e endereços dos serviços, foi utilizada a ferramenta DEByE (*Data Extraction By Example*) [LaRS02], desenvolvida no Laboratório de Bancos de Dados da UFMG, que permite especificar, por meio de exemplos, quais os dados que devem ser extraídos das páginas coletadas. Os dados extraídos em formato XML tiveram seus endereços geocodificados. Para isto, os endereços foram padronizados de forma a possibilitar sua geocodificação no banco de dados geográfico

[36] `http://www.belohorizonte.com.br (extinto)`
[37] `http://www.passeiolegal.com.br`
[38] `http://www.comidadibuteco.com.br`
[39] `http://www.terra.com.br`
[40] `http://www.inbh.com.br`
[41] Denomina-se de objeto o conjunto de dados de interesse que caracteriza cada um dos serviços descritos nas páginas coletadas (ex., para um restaurante, nome, endereço, categoria, etc.)

de Belo Horizonte. Eles foram transformados para o formato no qual os endereços estão armazenados no banco de dados geográfico, resultando em um conjunto estruturado de endereços que continha: (1) tipo de logradouro (rua, avenida, praça, etc.), (2) nome do logradouro, (3) número do imóvel e (4) bairro ou outros dados complementares.

Resumindo, o processo de geocodificação incluiu três etapas: (1) tratamento do endereço alfanumérico semi-estruturado extraído, (2) estabelecimento de uma correspondência entre o endereço extraído e o existente no banco de dados geográfico, e (3) atribuição de coordenadas para o endereço extraído. Naturalmente, a geocodificação é mais precisa quando o endereço extraído casa com um endereço individual existente no banco de dados geográfico. Quando não foi possível encontrar um endereço exato, a geocodificação foi feita pelo endereço numericamente mais próximo, no mesmo logradouro, ou considerando um local de referência. Por fim, os dados extraídos, com os endereços padronizados e coordenadas atribuídas (x,y), foram armazenados no banco de dados geográfico de Belo Horizonte.

Como resultado do experimento [BLMS03, LBCM05], foram criadas vinte e oito tabelas no banco de dados geográfico, uma para cada categoria de objeto extraído. Os atributos das tabelas criadas foram: nome do local, tipo de logradouro, nome do logradouro, número do imóvel, bairro e código do endereço. A Tabela 5.1 mostra o resultado da geocodificação. As colunas mostram o número de páginas coletadas, o número de objetos encontrados, o número de objetos extraídos, o número de objetos extraídos em que a localização exata pôde ser determinada, o número de objetos localizados em um endereço numericamente mais próximo ou em um local de referência e o número de endereços que não foram localizados. Como pode ser visto, 90% dos objetos foram localizados exatamente, 8% foram associados a um número aproximado ou a um ponto de referência e somente 2% não foram localizados devido ao endereço estar incompleto ou pertencer a cidades da região metropolitana de Belo Horizonte. Com isso foi possível visualizar em um mesmo ambiente geográfico todas os dados obtidos, de forma integrada. Na Figura 5.2, são visualizados, em um mesmo mapa, restaurantes, hotéis e um centro cultural, extraídos de diferentes *sites*. É possível, por exemplo, ver que o restaurante “Chez Bastião” fica ao lado do restaurante “Cafuso” e que ambos localizam-se perto do “Praça da Liberdade Hotel”. A Figura 5.2 também mostra que os mesmos dados podem ser visualizados sobre uma imagem de alta resolução de Belo Horizonte.

Tabela 5.1- Resultados Experimentais

Site	Páginas Coletadas	Objetos Encontrados	Objetos Extraídos	Geocodificação Exata	Geocodificação Aproximada	Não Geocodificado
Passeio Legal	12	122	122	120		2
Portal do Terra	2	52	52	48	4	
In BH	38	277	277	237	33	7
Comida di Buteco	13	89	89	83	6	

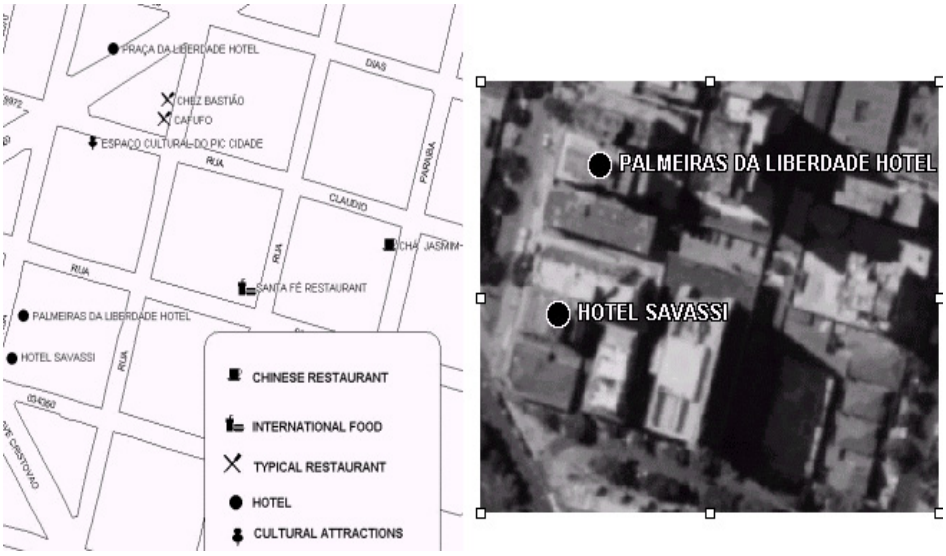


Figura 5.2 - Integração de dados sobre serviços extraídos de páginas da Web com os dados do banco de dados geográfico e imagens de alta resolução

Adicionalmente, foi possível visualizar todos os dados extraídos como uma aplicação na Web. A Figura 5.3 ilustra uma dessas páginas publicada utilizando o *software* Alov Map[42], onde os dados extraídos da Web são combinados com dados vindo do banco de dados geográfico de Belo Horizonte.

Com este experimento, percebemos que é possível extrair e geocodificar endereços presentes em páginas da Web gerando informação complementar, como as novas tabelas criadas no banco de dados geográfico. A extração e geocodificação foram consideradas satisfatórias. O próximo desafio seria então reconhecer e extrair os endereços de forma automática. Apesar do resultado satisfatório, o uso da ferramenta DEByE não é o mais adequado para uma extração automatizada de endereços. Além da DEByE necessitar que sejam fornecidos exemplos, ela depende da forma como a informação está estruturada na página ou seja, quanto mais homogênea for a página, em termos de marcação HTML ou de padrão, mais precisos serão os resultados. Devido às

[42] `http://alov.org/index.htm`

suas características, um endereço pode ser descrito de diversas maneiras em páginas da Web, aparecendo muitas vezes sem uma marcação bem definida, como, por exemplo, no meio de textos, o que não facilita o seu reconhecimento e extração pela DEByE. Conforme será visto na próxima seção, para realizar tal tarefa de forma mais eficiente foi necessário usar a ontologia de extração proposta nesta tese.

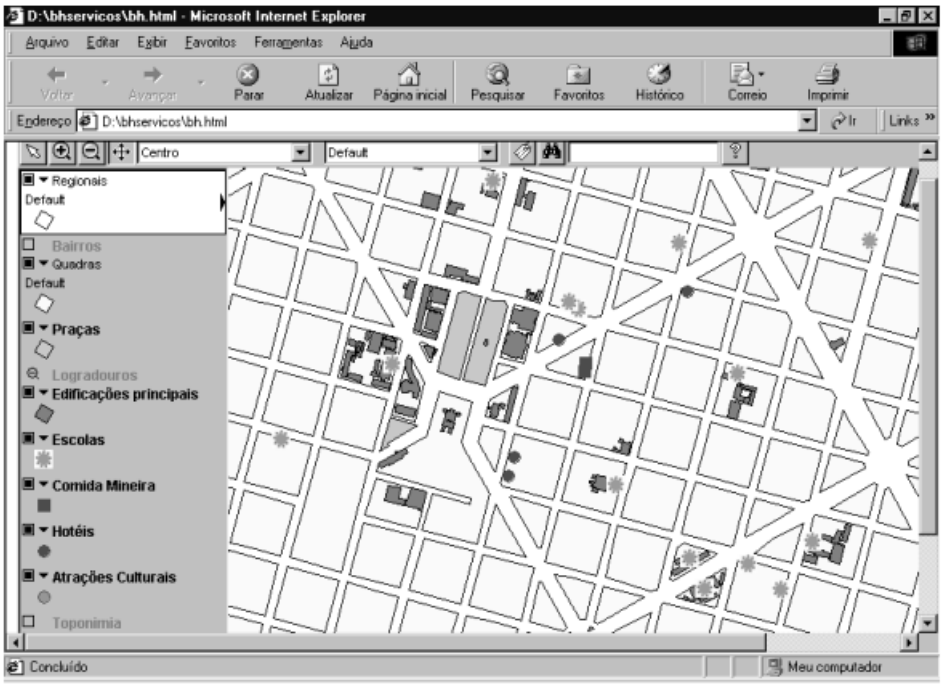


Figura 5.3 – Mapa com restaurantes de comida mineira, hotéis e atrações culturais extraídas de páginas da Web e escolas extraídas do banco de dados geográfico de Belo Horizonte

Este experimento também mostrou a viabilidade de utilização de dados sobre serviços e atividades, extraídos de páginas da Web, por ferramentas como *Google Earth* e *Virtual Earth*, tendo como base as coordenadas geográficas dos endereços extraídos. Da mesma forma que os dados sobre serviços extraídos foram visualizados sobre imagens de Belo Horizonte, eles poderiam ser visualizados sobre as imagens de satélite providas pelo *Google Earth*. A Figura 5.4 mostra a região do Parque Municipal em Belo Horizonte visualizada no *Google Earth*. Como pode ser percebido, não existe nenhuma informação agregada. Já na Figura 5.5, a mesma área do Parque Municipal é visualizada como resultado deste experimento, apresentando a localização de hotéis, restaurantes e centros culturais cujos dados foram extraídos das páginas da Web utilizadas. Usando as coordenadas geográficas dos endereços, este resultado poderia ser visualizado no *Google Earth* permitindo consultas e serviços de busca local.

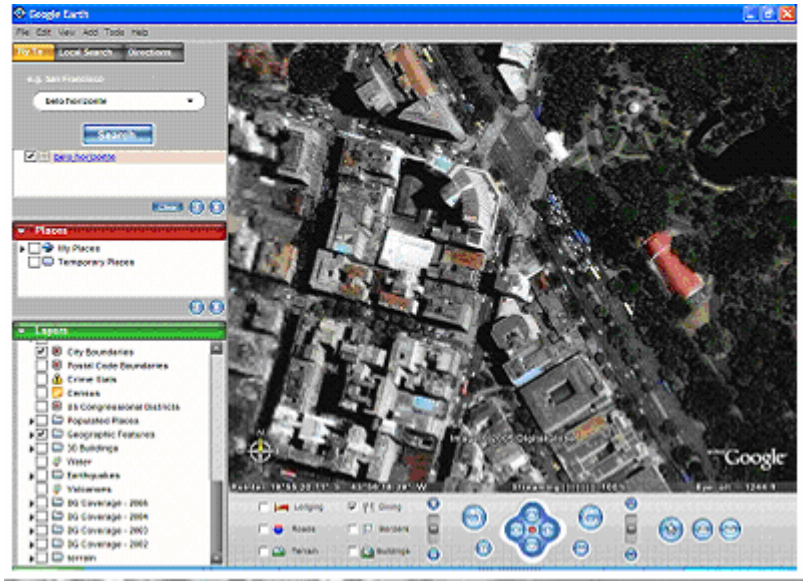


Figura 5.4 – Visualização no Google Earth da área próxima ao Parque Municipal em Belo Horizonte.

Figura 5.5 – Visualização em um SIG da área próxima ao Parque Municipal em Belo Horizonte com a localização de restaurantes, hotéis e centros culturais extraídos de páginas da Web

É importante ressaltar, ainda, que a abordagem utilizada neste experimento foi posteriormente utilizada por Souza [Souz05] para alimentar o *gazetteer* do *Locus* com lugares intra-urbanos como pontos turísticos e serviços (hotéis, restaurantes, teatros, museus e outros) cujos dados foram extraídos de páginas da Web brasileira conforme relatado na Seção 3.5.3.

5.2 Experimento 2 – Avaliação dos Padrões para Extração Automática de Endereços de Páginas da Web

Dando seqüência ao experimento anterior, este experimento teve por objetivo principal verificar a eficácia da nossa abordagem para extrair automaticamente endereços de páginas da Web. Para isto, foram utilizados os padrões para reconhecimento de endereços em páginas da Web definidos no Capítulo 4. Uma vez definidos esses padrões, era necessário conhecer como os respectivos endereços estão presentes na Web brasileira e avaliar a sua capacidade de extração.

O experimento descrito nesta seção procurou responder às seguintes perguntas: (1) Qual o percentual de ocorrências de endereços nas páginas? (2) Qual o percentual de ocorrência de cada padrão? (3) Qual a forma mais comum de endereço encontrada? (4) Qual o padrão mais confiável? (5) Qual o melhor identificador de localização para geocodificação? (6) Qual a qualidade dos endereços extraídos? Com a realização deste experimento, foi possível definir um conjunto mínimo de padrões que fosse eficaz e confiável para a extração e geocodificação de endereços.

Optou-se por usar como coleção de teste para este experimento, a mesma coleção utilizada em [Delb05] para identificação e classificação de expressões de posicionamento em páginas da Web, devido às suas características. Ela foi obtida através da submissão automática ao Google dos nomes de um conjunto representativo de pontos de referência de Belo Horizonte bem como da denominação de alguns pontos de referência genéricos como "aeroporto", "centro" e "rodoviária". Todas as páginas retornadas como resposta foram coletadas e posteriormente pré-processadas. Em uma coleção como essa, criada a partir de nomes de pontos de referência, existe a probabilidade de se encontrar um volume representativo de páginas com a presença de evidências geo-espaciais sob a forma de endereços, sendo, portanto, de grande valia para o tipo de avaliação que o experimento se propõe.

O experimento foi dividido em cinco etapas: (1) extração de endereços de uma coleção com 75.413 páginas da Web brasileira usando todos os padrões definidos; (2) validação dos endereços extraídos usando a geocodificação dos identificadores de lugar; (3) validação do conteúdo extraído usando uma amostragem de 385 extrações para cada

padrão definido; (4) validação dos endereços extraídos usando a geocodificação dos endereços; (5) avaliação geral da capacidade de extração usando uma amostragem de 385 páginas. As cinco etapas do experimento serviram para ajustar os padrões inicialmente definidos, avaliar a eficácia dos padrões para extração automática de endereços e selecionar os seis padrões mais eficazes e confiáveis. Na coleção utilizada, os marcadores HTML foram substituídos pelo separador <S> para simplificar o processamento das páginas.

5.2.1 Reconhecimento e Extração de Endereços

Com base nos padrões definidos, foi desenvolvido um extrator na linguagem PERL, que executa o *geoparsing* do endereço e sua extração. O extrator procura por cada um dos dezoito padrões que representam as variações mais comuns encontradas em endereços na Web. As expressões regulares definidas para cada padrão e utilizadas pelo extrator estão detalhadas no Capítulo 4 e a sua versão em PERL encontra-se no Anexo C. Os dezoito padrões estão listados abaixo:

1. *Telefone*
2. *CidadeEstado* + *Telefone*
3. *Telefone* + *CidadeEstado*
4. *CEP* + *Telefone*
5. *Telefone* + *CEP*
6. *Endereço Básico*+ *CEP* + *CidadeEstado* + *Telefone*
7. *Endereço Básico* + *CidadeEstado* + *CEP* + *Telefone*
8. *Endereço Básico* + *CEP* + *CidadeEstado*
9. *Endereço Básico* + *CidadeEstado* + *CEP*
10. *Endereço Básico* + *Telefone*
11. *Endereço Básico* + *Telefone* + *CidadeEstado*
12. *Endereço Básico* + *CidadeEstado* + *Telefone*
13. *Endereço Básico* + *CidadeEstado*
14. *Endereço Básico* + *CEP* + *Telefone*
15. *Endereço Básico* + *Telefone* + *CEP*
16. *Endereço Básico* + *CEP*
17. *CEP*
18. *CidadeEstado*

A coleção de 75.413 páginas foi submetida ao extrator, sendo os endereços localizados nas páginas por meio de casamento de padrão.

5.2.2 Funcionamento do Extrator

O procedimento geral do extrator pode ser descrito em três passos: (1) cada página coletada é percorrida à procura de cada um dos dezoito padrões de endereço; (2) se algum casamento de padrão ocorre, um endereço em potencial é reconhecido; (3) os termos correspondentes são extraídos e armazenados em um repositório temporário para posterior validação.

Os termos extraídos da coleção foram armazenados no repositório temporário considerando-se os seguintes campos: URL, posição inicial do casamento, posição final do casamento, tipo do padrão de extração (1 a 18), tipo de logradouro, nome do logradouro, número do imóvel, complemento, DDD, prefixo do telefone, sufixo do telefone, CEP, cidade, estado. O nome da URL e a posição (início e fim), encontrados no texto foram preenchidos para todos os padrões. Já os demais campos, foram preenchidos de acordo com o tipo de padrão encontrado.

O complemento de um endereço é um complicador no processo de reconhecimento de padrões de endereço pelo fato de ser opcional, ocupar um número indefinido de posições e possuir um conteúdo bem diversificado. Muitas vezes não existe separador entre o complemento e o nome da cidade, não sendo, assim, possível separá-lo corretamente como, por exemplo, no endereço "R Dom Silverio, 461 <S> ; Passos Juiz de Fora - MG - CEP: 36026-450". Outras vezes, o complemento é muito longo, dificultando o seu reconhecimento, como no caso "<S>Rua Rodrigues Caldas, 79 Edificio Tiradentes - 22 andar - sala 1<S> Bairro: Santo Agostinho<S> Belo Horizonte - MG <S> CEP 30190-921<S>".

Para tratar a opcionalidade do complemento e a questão do seu tamanho, o extrator adota uma estratégia simples e considera como complemento os componentes de endereço encontrados entre um número de imóvel e o primeiro identificador de localização (*Telefone*, *CEP* ou *CidadeEstado*) que existir no endereço encontrado. O tamanho do complemento também foi limitado a 30 posições, que é um tamanho médio, escolhido heuristicamente depois de várias tentativas. É importante ressaltar que, com um limite maior, a probabilidade de ocorrer um erro de reconhecimento aumenta

consideravelmente. As Figuras 5.6 e 5.7 ilustram o funcionamento do extrator através de um exemplo. Na Figura 5.6 encontra-se o endereço que será extraído.

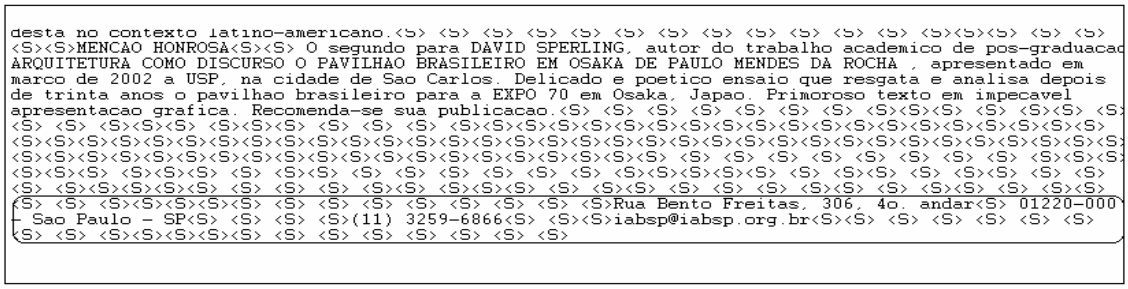


Figura 5.6 – Presença de um endereço no código HTML de uma página da Web

O endereço "Rua Bento de Freitas, 306, 4o. andar 012220-000 - São Paulo – SP (11) 3259-6866 " foi tratado pelos seguintes padrões: *Telefone* (1), *CidadeEstado* + *Telefone* (2), *Endereço Básico* + *CEP* + *CidadeEstado* + *Telefone* (6), *Endereço Básico* + *CidadeEstado* + *CEP* (8), *Endereço Básico* + *CidadeEstado* + *Telefone* (12), *Endereço Básico* + *CidadeEstado* (13), *Endereço Básico* + *CEP* (16), *CEP* (17) e *CidadeEstado* (18). O resultado de cada extração, correspondente a cada padrão aplicado, é mostrado na Figura 5.7.

Conforme observado na Figura 5.7, o uso dos vários padrões gera repetições na identificação de um mesmo endereço. Para eliminar as duplicações é executada uma rotina que identifica o padrão mais abrangente e, com isto, um único padrão é considerado. A rotina que elimina as duplicações compara todas as ocorrências encontradas em uma mesma página, que estejam contidas em um determinado intervalo - posição de início e fim, e é denominada, nesta tese, de *normalização*. O padrão mais abrangente tem o intervalo maior e engloba as informações existentes nos outros. No exemplo da Figura 5.7, após a normalização, ficaria apenas o endereço reconhecido pelo padrão *Endereço Básico* + *CEP* + *CidadeEstado* + *Telefone* (6).

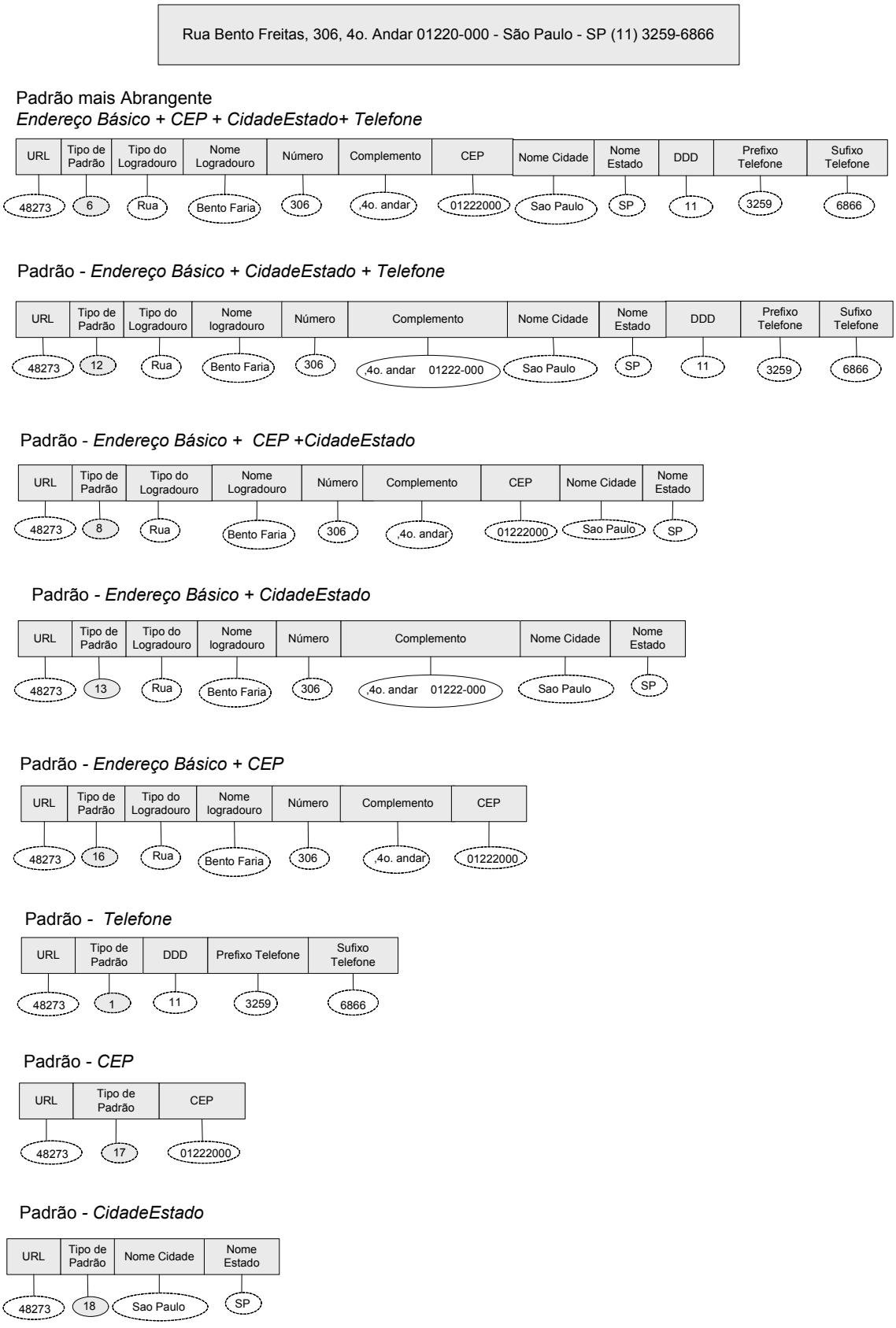


Figura 5.7 – Extração do endereço considerando diversos padrões

5.2.3 Resultado da Extração

Após a extração ter sido executada nas 75.413 páginas da coleção, foram encontrados 893.260 endereços em 43.121 páginas, ou seja, 57% das páginas continham endereços formatados de acordo com algum tipo de padrão (Tabela 5.2). O alto percentual de páginas com endereços reconhecidos se deve as características da coleção utilizada, o que confirma e valida os motivos da escolha dessa coleção para a execução do experimento. A Tabela 5.3 apresenta, em ordem crescente de número de extrações, o total de endereços extraídos com cada um dos dezoito padrões. A validação da extração encontra-se detalhada nas seções 5.2.4 e 5.2.5.

Tabela 5.2 – Resumo da extração de endereços

Total de páginas	75.413
Total de páginas com endereços reconhecidos	43.121
Total de endereços reconhecidos	893.260

Avaliando os resultados obtidos, observa-se que a ocorrência de endereços parciais (*CEP*, *Telefone*, *CidadeEstado*) foi maior, e que um *Endereço Básico* seguido de apenas um dos identificadores ocorre com mais freqüência do que um endereço na forma mais completa. Os resultados foram bastante satisfatórios mostrando que é possível reconhecer e extrair automaticamente endereços de páginas da Web. Além disso, os resultados indicam que mesmo quando não é feito o reconhecimento de um endereço no formato convencional, é possível reconhecer e utilizar o endereço parcial, garantindo que pelo menos uma cidade e um estado possam ser associados a uma determinada porção da página.

Esse experimento possibilitou uma visão geral de como os endereços são estruturados e, como pode ser percebido, mostrou que não é necessário um padrão tão completo para se identificar um lugar com base em seu endereço. Padrões menores estão, geralmente, contidos em padrões maiores e recuperam os mesmos endereços, apenas com menos detalhes. Observando a Figura 5.8 percebemos que o *Endereço Básico* acompanhado do *CEP* é recuperado por qualquer um dos padrões (7, 8, 14, 16) exemplificados na figura ou seja, se objetivo for reconhecimento de um endereço e de sua localização não seria necessário procurar por um padrão de endereço como os de

número 6, 7 ou 9. Os padrões de número 6 e 9 também seria reconhecidos pelo padrão *Endereço Básico + CEP* desde que *CidadeEstad*o ocupasse menos de 30 posições.

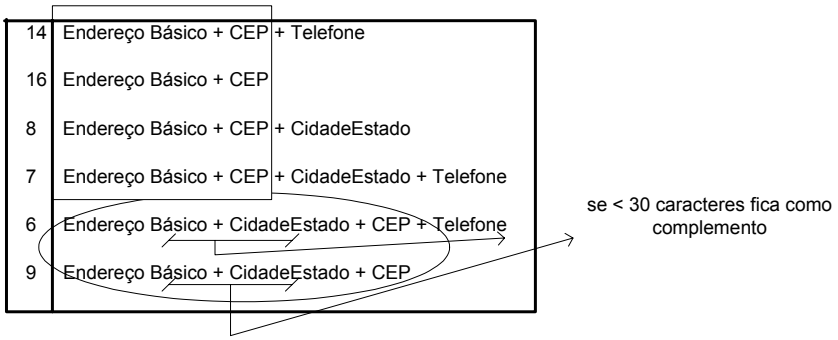


Figura 5.8 – Inter-relação entre os padrões para extração de endereços

Tabela 5.3 – Resultado da extração de endereços por tipo de padrão

ID	Tipo de Padrão	Total de Extrações Executadas
7	*Endereço Básico + CidadeEstado + CEP + Telefone*	319
6	*Endereço Básico + CEP + CidadeEstado + Telefone*	383
2	*CidadeEstado + Telefone*	433
15	*Endereço Básico + Telefone + CEP*	449
12	*Endereço Básico + CidadeEstado + Telefone*	525
11	*Endereço Básico + Telefone + CidadeEstado*	625
5	*Telefone + CEP*	658
3	*Telefone + CidadeEstado*	858
14	*Endereço Básico + CEP + Telefone*	1.860
4	*CEP + Telefone*	2.358
9	*Endereço Básico + CidadeEstado + CEP*	6.981
13	*Endereço Básico + CidadeEstado*	19.067
8	*Endereço Básico + CEP + CidadeEstado*	21.424
10	*Endereço Básico + Telefone*	26.350
16	*Endereço Básico + CEP*	43.074
17	*CEP*	99.842
1	*Telefone*	151.641
18	*CidadeEstado*	516.413
	Total	**893.260**

Sendo os endereços identificados e extraídos, surge a necessidade de validar a extração. A primeira estratégia adotada foi geocodificar os identificadores de localização, verificando no *Locus* a existência de cada um deles. Havendo casamento do identificador de localização contido no endereço extraído com algum CEP, telefone ou cidade/estado armazenado no *gazetteer* do *Locus*, é grande a probabilidade de a extração ter sido correta (Seção 5.2.4). Sem uma correspondência no *Locus* não é possível validar de onde é o endereço. A segunda estratégia foi selecionar, dentro de uma amostragem, todas as extrações que não foram geocodificadas pelos seus identificadores de localização e verificar o conteúdo extraído (Seção 5.2.5.1). A terceira estratégia foi verificar, nas extrações geocodificadas, se a parte correspondente ao *Endereço Básico* foi extraída corretamente, adotando-se a geocodificação de endereços (Seção 5.2.5.2). As próximas seções discutem as estratégias de validação da extração realizada.

5.2.4 Validação dos Endereços Extraídos: Geocodificação dos Identificadores de Localização

Com a geocodificação dos identificadores de localização, encontrados nos endereços extraídos, era esperado: (1) validar a extração através da existência do lugar correspondente no *gazetteer* do *Locus*, usando para isto, o número de telefone identificado, o CEP ou o par CidadeEstado; (2) verificar se, nos endereços que possuem mais de um identificador, existe incompatibilidade entre eles; (3) verificar qual é o identificador de localização mais confiável para geocodificação; e (4) verificar se são necessários os dezoito padrões para reconhecer um endereço.

Essa estratégia de geocodificação dos identificadores de localização teve por objetivo suprir a falta de uma coleção de teste onde fosse possível comparar se toda a extração havia sido realizada corretamente. Devido ao volume de dados manipulados, essa estratégia permitiu filtrar os endereços para os quais não se conseguiu achar um lugar correspondente no *Locus*, considerando corretos aqueles para os quais tal correspondência foi encontrada.

O procedimento adotado para a geocodificação dos identificadores de localização é descrito a seguir. Para cada endereço extraído, o identificador de localização encontrado foi submetido ao *Locus*. Quando esse identificador era o *Telefone*, foi consultado se o "DDD + prefixo" existiam e, quando possível, eram recuperados o nome da cidade e do

estado correspondentes. O mesmo procedimento foi feito para o *CEP* e para os nomes de cidade e estado extraídos. Entretanto, existindo uma ocorrência de *Telefone*, a geocodificação foi primeiro executada por este identificador, tendo em vista a sua presença em quase todos os padrões e a sua alta incidência na Web (ver Tabela 5.3). Assim, a seqüência em que os identificadores de localização foram verificados foi: primeiro *Telefone*, depois *CEP* e por fim *CidadeEstado*, que por ser um identificador textual, tem mais possibilidade de apresentar erros. Finalmente, cada geocodificação foi comparada com as demais já obtidas para fins de verificação de compatibilidade. Por exemplo, feita a geocodificação por *Telefone* e logo após pelo *CEP*, a cidade e o estado atribuídos pelo *CEP* foram comparados com a cidade e o estado atribuídos pelo *Telefone*.

Com este procedimento, foi possível avaliar a qualidade das informações de lugar fornecidas para cada endereço extraído, bem como qual identificador apresentou menos erros, obtendo um maior percentual de geocodificação. Para os casos onde não foi possível encontrar correspondência no *Locus*, não foram atribuídas coordenadas do RME do município.

A Tabela 5.4 apresenta o resumo dos resultados obtidos com o procedimento de geocodificação executado. Para cada padrão é apresentado: (a) a quantidade total de endereços extraídos; (b) a quantidade e o percentual de endereços geocodificados usando o *Telefone*; (c) a quantidade e o percentual de endereços geocodificados usando o *CEP* e o percentual de geocodificação diferente do obtido com o *Telefone*; (d) a quantidade e o percentual de endereços geocodificados usando *CidadeEstado* e o percentual de geocodificação diferente do obtido com o *Telefone* e com o *CEP*. Por exemplo, usando-se o padrão (4) *CEP* + *Telefone*, foram extraídos 2.358 endereços, dos quais 1.611 foram geocodificados pelo *Telefone* e 2.284 pelo *CEP*. Isto significa que um número maior de CEP foi encontrado no *Locus*. Dentre as 1611 extrações em comum, apenas 2,76 % identificaram locais diferentes.

Analisando os resultados da Tabela 5.4 é possível verificar o grau de discrepância entre as geocodificações e avaliar a eficácia de cada um dos identificadores de localização na geocodificação. Observa-se que, de uma maneira geral, o *CEP* é mais eficaz para a geocodificação quando comparado aos demais identificadores, conseguindo sempre uma geocodificação em torno de 90%. A geocodificação através do

número do telefone foi em média 67%. Apesar dos problemas de grafia e abreviaturas e do procedimento só considerar o casamento exato, o identificador *CidadeEstado* também foi bastante eficaz na geocodificação. No entanto, o uso deste identificador isoladamente, como no padrão (18), não é recomendado, mesmo tendo sido grande o número de ocorrências reconhecidas. Através de uma inspeção visual, foi detectado que o percentual de apenas 50,76% de geocodificação foi devido a enganos no reconhecimento do nome da cidade, principalmente nos casos onde os nomes da cidade e do estado não vêm seguidos ou são precedidos de outros componentes do endereço. Isso porque o padrão procura por um nome iniciado por letra maiúscula precedido por um nome ou sigla de estado e, como exemplificado na Figura 5.9, mesmo fixando os valores para nome de estado, o que deveria ser o nome da cidade aparece em diferentes contextos. A geocodificação serve, portanto, como um filtro eliminando os casos em que o reconhecimento foi equivocado. Por outro lado, esse padrão pode ser utilizado para reconhecimento de alguns lugares de referência associados a uma lista de palavras claramente reconhecidas como lugar. Por exemplo, foram encontradas muitas ocorrências com as palavras "universidade", "campus", "cine", "museu", "parque", "parque nacional", "palácio", etc. , que denominam lugares de interesse, conforme pode ser observado na Figura 5.9.

Nome Cidade	**Nome Estado**
Abrasel	BA
Academia Belo Horizonte	MG
Academia Brasileira de Letras	Rio de Janeiro
Abril Cultural	São Paulo
Aeroclube de Manaus	AM
Cine Eldorado	SP
Univap Sao Jose dos Campos	SP

Figura 5.9 – Exemplos de problemas no uso do padrão *CidadeEstado*

A discrepância entre as geocodificações utilizando os identificadores *CEP*, *Telefone* e *CidadeEstado* foram relativamente baixas, ficando sempre abaixo de 9% e atingindo uma média de 2,4% entre *CEP* e *Telefone*, 4,4% entre *CidadeEstado* e *Telefone*, e

2,22% entre *CidadeEstado* e *CEP*. A maior diferença ocorreu entre a geocodificação usando os identificadores *Telefone* e *CidadeEstado*. Os motivos que levaram a essas diferenças são discutidos na Seção 5.2.5.1.

Os resultados confirmaram que, para geocodificar, o melhor identificador de localização é o *CEP*, seguido pelo *Telefone* e, por último, por *CidadeEstado* sempre que associado com algum outro componente de endereçamento. O padrão 18 (*CidadeEstado*), entretanto, não deve ser usado devido aos motivos já mencionados. O *Telefone*, por ser um padrão numérico e estar presente em quase todos os demais padrões definidos, possibilita um reconhecimento menos propenso a erros e mais simples de ser efetuado.

Como o objetivo é obter o menor número possível de padrões a serem utilizados na prática, principalmente no caso de coleções grandes, os dezoito padrões inicialmente definidos foram reduzidos a onze. A Tabela 5.5 apresenta os onze padrões que obtiveram um percentual maior no reconhecimento de endereços: *Telefone* (1), *Telefone* + *CidadeEstado* (3), *CEP* + *Telefone* (4), *Endereço Básico* + *CEP* + *CidadeEstado* (8), *Endereço Básico* + *CidadeEstado* + *CEP* (9*), Endereço Básico* + *Telefone* (10), *Endereço Básico* + *CidadeEstado* (13), *Endereço Básico* + *CEP* + *Telefone* (14), *Endereço Básico* + *CEP (16)*, *CEP* (17) e *CidadeEstado* (18).

Eliminado o padrão *CidadeEstado* (18) por não ser considerado um padrão eficaz, o conjunto passou a ter dez padrões. Analisando esses dez padrões, foram eliminados os que continham maior número de componentes e que poderiam ser substituídos por padrões mais eficazes e com menor número de componentes, como exemplificado na Figura 5.8. Como resultado dessa análise, seis padrões foram selecionados (Tabela 5.6). A escolha desses seis padrões privilegiou aqueles que possuíam os identificadores de localização *CEP* e *Telefone*. O único padrão com maior número de componentes mantido foi *Endereço Básico* + *CidadeEstado* + *CEP* (9), que obteve um percentual bastante representativo de endereços extraídos e possibilita uma geocodificação mais eficiente devido à presença do CEP.

Tabela 5.4 - Total de endereços geocodificados

ID	Tipo de Padrão	Total de Endereços Extraídos	*Telefone*	*CEP*		*CidadeEstado*		
			Qte Geo-codificada	Qte Geo-codificada	% Geo-codificado diferente Telefone	Qte Geo-codificada	% Geo-codificado diferente Telefone	% Geoco-dificado diferente CEP
1	*Telefone*	151.641	105.258 69,41%					
2	*CidadeEstado + Telefone*	433	338 78,06%			290 66,97%	5,86%	
3	*Telefone + CidadeEstado*	858	598 69,70%			751 87,53%	4,53%	
4	*CEP + Telefone*	2.358	1.611 68,32%	2.284 96,86%	2,76%			
5	*Telefone + CEP*	658	429 65,20%	591 89,82%	3,89%			
6	*Endereço Básico + CEP + CidadeEstado + Telefone*	383	239 62,40%	379 98,96%	0,79%	377 98,43%	1,59%	2,39%
7	*Endereço Básico + CidadeEstado + CEP + Telefone*	319	220 68,97%	318 99,69%	1,89%	283 88,71%	1,41%	2,12%
8	*Endereço Básico + CEP + CidadeEstado*	21.424		21.219 99,04%		20.436 95,39%		2,06%
9	*Endereço Básico + CidadeEstado + CEP*	6.981		6.905 98,91%		4.207 60,26%		2,31%
10	*Endereço Básico + Telefone*	26.350	17.245 65,45%					
11	*Endereço Básico + Telefone + CidadeEstado*	625	419 67,04%			521 83,36%	8,64%	
12	*Endereco Básico + CidadeEstado + Telefone*	525	447 85,14%			426 81,14%	2,82%	
13	*Endereço Básico + CidadeEstado*	19.067				13.255 69,52%		
14	*Endereço Básico + CEP + Telefone*	1.860	1193 64,14%	1.805 97,04%	3,71%			
15	*Endereço Básico + Telefone + CEP*	449	186 41,43%	439 97,77%	1,59%			
16	*Endereço Básico + CEP*	43.074		42.327 98,27%				
17	*CEP*	99.842		86.730 86,87%				
18	*CidadeEstado*	516.413				262.108 50,76%		

Tabela 5.5 – Total de endereços extraídos e geocodificados com os onze padrões

ID	Tipo de Padrão	Total Extraído	Total Geocodificado
1	*Telefone*	151.641	105.258 69,41%
3	*Telefone + CidadeEstado*	858	751 87,53%
4	*CEP + Telefone*	2.358	2.284 96,86%
8	*Endereço Básico + CEP + CidadeEstado*	21.424	21.219 99,04%
9	*Endereço Básico + CidadeEstado + CEP*	6.981	6.905 98,91%
10	*Endereço Básico + Telefone*	26.350	17.245 65,45%
13	*Endereço Básico + CidadeEstado*	19.067	13.255 69,52%
14	*Endereço Básico + CEP + Telefone*	1.860	1.805 97,04%
16	*Endereço Básico + CEP*	43.074	42.327 98,27%
17	*CEP*	99.842	86.730 86,87%
18	*CidadeEstado*	516.413	262.108 50,76%

Tabela 5.6 – Padrões selecionados para reconhecimento de endereços

ID	Padrões Selecionados
1	*Telefone*
9	*Endereço Básico + CidadeEstado + CEP*
10	*Endereço Básico + Telefone*
13	*Endereço Básico + CidadeEstado*
16	*Endereço Básico + CEP*
17	*CEP*

5.2.5 Avaliação dos Padrões de Extração

A avaliação dos padrões de extração foi dividida em duas etapas: (1) avaliação do conteúdo (O que foi extraído estava correto?) e (2) avaliação da capacidade de extração (Quanto, do que deveria ser extraído, foi realmente extraído?). Essas duas etapas avaliam a eficácia do resultado da extração e correspondem, respectivamente, a duas medidas de avaliação amplamente usadas na área de RI: precisão[43] e revocação[44] [BYRN99]. Portanto, o objetivo desta avaliação foi aferir a qualidade da extração e

[43] Precisão é a fração de documentos recuperados que são relevantes
[44] Revocação é a fração de documentos relevantes de uma coleção que foram recuperados

confirmar se realmente os seis padrões selecionados seriam os mais abrangentes e eficazes.

Para essa avaliação, foi utilizado um espaço amostral de 385 páginas, cujo tamanho, *n*, foi obtido pela fórmula de amostra aleatória simples (1), onde N representa o total de páginas, E_0^2 o erro amostral tolerável (5%) e n_0 a primeira aproximação do tamanho da amostra [Barb99].

$$n = \frac{N \times n_0}{N + n_0}, \quad \text{onde} \quad n_0 = \frac{1}{E_0^{\,2}} \qquad (1)$$

Na avaliação, foi verificado o conteúdo do identificador de localização nas extrações que não foram geocodificadas (Seção 5.2.5.1), analisando-se o motivo da não geocodificação, e o conteúdo dos componentes do *Endereço Básico* nas extrações geocodificadas, analisando-se a qualidade de geocodificação desses endereços (Seção 5.2.5.2). Para avaliar a capacidade do extrator, foi conferido se o extrator conseguiu reconhecer todos os endereços existentes nas páginas (Seção 5.2.5.3).

5.2.5.1 Avaliação do Conteúdo – Identificadores de Localização

Para avaliar a eficácia de extração de cada um dos padrões, foi considerado a princípio, um espaço amostral de 385 páginas. No entanto, dentro desse espaço amostral, alguns padrões tiveram uma ocorrência muito pequena (Tabela 5.7). Para obter uma maior representatividade, optou-se por aplicar a fórmula de amostra aleatória simples no total de extrações obtidas com cada padrão (Tabela 5.5), comparando os valores obtidos. Doze padrões tiveram seu espaço amostral entre 175 e 378. Os seis padrões de maior ocorrência tiveram seu espaço amostral entre 393 e 398. Com base nesses valores, fixamos para esta avaliação, um número de 385 extrações para cada tipo de padrão por considerar um valor bastante representativo dentro do espaço amostral das 385 páginas. Esse número contempla todas as ocorrências dos padrões de menor incidência e aproxima-se muito do tamanho do espaço amostral daqueles com maior ocorrência, considerando o total de extrações obtidas.

Tabela 5.7 – Total das extrações encontradas na amostragem

ID	Tipo de Padrão	Total de Extrações
1	*Telefone*	2.818
2	*CidadeEstado + Telefone*	36
3	*Telefone + CidadeEstado*	10
4	*CEP + Telefone*	80
5	*Telefone + CEP*	6
6	*Endereço Básico + CEP + CidadeEstado + Telefone*	1
7	*Endereço Básico + CidadeEstado + CEP + Telefone*	0
8	*Endereço Básico + CEP + CidadeEstado*	353
9	*Endereço Básico + CidadeEstado + CEP*	419
10	*Endereço Básico + Telefone*	518
11	*Endereço Básico + Telefone + CidadeEstado*	7
12	*Endereço Básico + CidadeEstado + Telefone*	29
13	*Endereço Básico + CidadeEstado*	781
14	*Endereço Básico + CEP + Telefone*	59
15	*Endereço Básico + Telefone + CEP*	3
16	*Endereço Básico + CEP*	979
17	*CEP*	2970
	Total	9069

Portanto, foram escolhidas aleatoriamente 385 extrações de cada tipo de padrão, exceto para o padrão *CidadeEstado* (18), que devido aos problemas discutidos na seção anterior, não foi considerado. Para aqueles padrões onde não foi possível encontrar essa quantidade de extrações dentro da amostragem, o restante foi obtido aleatoriamente dentro do total de endereços extraídos. Os padrões 6 e 7 que tiveram menos de 385 ocorrências ficaram com o número total de extrações. Assim, para esta avaliação foram consideradas 6.475 ocorrências de endereços distribuídas conforme a Tabela 5. 8.

Dentre as 6.475 ocorrências, apenas 549 não foram geocodificadas, ou seja, 8,5% das extrações não tiveram o nome da cidade e do estado atribuídos pelo *Locus* e, portanto, ficaram sem coordenadas geográficas associadas. Para verificar porque não houve a geocodificação, cada uma das 549 extrações foi analisada manualmente, avaliando-se se os identificadores de localização extraídos estavam corretos ou não. A Tabela 5.8 apresenta também a distribuição das 549 ocorrências por tipo de padrão,

destacando, em negrito, os seis padrões selecionados na seção anterior como sendo os mais eficazes para a extração de endereços de páginas da Web. A Tabela 5.8 também identifica nove padrões que produziram extrações totalmente corretas. Esses padrões, marcados com um asterisco, possibilitaram a extração de endereços que foram totalmente geocodificados e, os que não foram geocodificados, não o foram devido ao conteúdo informado e não à extração efetuada (ver Tabela 5.9).

Tabela 5.8 – Padrões com resultado da geocodificação

ID	Tipo de Padrão	Qte total	Qte sem geocodificação	
1	***Telefone***	**385**	**123**	**31,9%**
2	*CidadeEstado + Telefone*	385	24	6,2%
3	*Telefone + CidadeEstado*	385	14	3,6%
4	*CEP + Telefone* **(*)**	385	3	0,8%
5	*Telefone + CEP* **(*)**	385	9	2,3%
6	*Endereço Básico + CEP + CidadeEstado + Telefone* **(*)**	382	0	0%
7	*Endereço Básico + CidadeEstado + CEP + Telefone* ***(*)***	318	0	0%
8	*Endereço Básico + CEP + CidadeEstado* **(*)**	385	2	0,5%
9	***Endereço Básico + CidadeEstado + CEP*** **(*)**	**385**	**0**	**0%**
10	***Endereço Básico + Telefone***	**385**	**116**	**30,1%**
11	*Endereço Básico + Telefone + CidadeEstado*	385	21	5,45%
12	*Endereço Básico + CidadeEstado + Telefone*	385	4	1,03%
13	***Endereço Básico + CidadeEstado***	**385**	**192**	**49,9%**
14	*Endereço Básico + CEP + Telefone* **(*)**	385	4	1,04%
15	*Endereço Básico + Telefone + CEP* **(*)**	385	6	1,6%
16	***Endereço Básico + CEP*** **(*)**	**385**	**5**	**1,3%**
17	***CEP***	**385**	**26**	**6,8%**
	TOTAL	**6475**	**549**	**8,5%**

É interessante observar que as ocorrências que foram totalmente geocodificadas seguem padrões de endereço mais completos e, mesmo assim, a extração foi bem sucedida. O pior percentual de geocodificação, 49,9%, ficou com o padrão *Endereço Básico + CidadeEstado* (13), seguido pelos padrões *Telefone* (1) e *Endereço Básico + Telefone* (10). Como a representatividade desses padrões é alta, estando eles entre os seis selecionados, foi necessário investigar se o problema estava na extração ou na geocodificação.

Analisando a Tabela 5.9, conclui-se que para os padrões *Telefone* (1) e *Endereço Básico* + *Telefone* (10), o problema estava no conteúdo informado e não no padrão de extração. Ambos apresentaram um percentual baixo de extração errada. No entanto, para o padrão *Endereço Básico* + *CidadeEstado* (13), o problema se deu no reconhecimento do nome da cidade. Uma discussão mais detalhada dos problemas encontrados é feita no decorrer desta seção.

A Tabela 5.9 apresenta, para os padrões com ocorrências sem geocodificação, o resultado detalhado indicando, na coluna "Total", a quantidade de extrações que não foram geocodificadas, na coluna "Extrações Corretas", a quantidade de extrações que não puderam ser geocodificadas mas foram extraídas corretamente, e na coluna "Extrações com erro", a quantidade de extrações com identificadores de localização errados. Os seis padrões, em negrito, tiveram extrações consideradas corretas, porém elas não foram geocodificadas por não encontrarem uma correspondência de CEP ou número de telefone no *Locus*. Os motivos que impediram essa geocodificação estão relacionados ao número de telefone e CEP informados foram os seguintes:

1. Número de Telefone
   - números desatualizados não possuindo quatro dígitos no prefixo;
   - números com código de área errado;
   - números não encontrados no *Locus* – as informações de telefone foram obtidas do site da Anatel e podem estar desatualizadas.
2. CEP
   - códigos errados;
   - códigos não encontrados no *Locus* – as informações de CEP utilizadas são referentes a 2000, podendo estar desatualizadas.

Os oito padrões restantes apresentaram algum tipo de erro na extração. Analisando a Tabela 5.9, verificou-se que as extrações erradas concentraram-se principalmente nas extrações cujos padrões possuem, como um de seus dos componentes, os identificadores de localização *CidadeEstado* ou *CEP* (padrões 11, 13 e 17). A inspeção visual do resultado da extração identificou o que ocorreu em cada um deles.

Tabela 5.9 – Detalhamento das extrações sem geocodificação

ID	Tipo de Padrão	Extrações sem Geocodificação				
		Total	Extrações corretas		Extrações com erro	
1	*Telefone*	123	120	97,6%	3	2,4%
2	*CidadeEstado + Telefone*	24	4	16,7%	20	83,3%
3	*Telefone + CidadeEstado*	14	0	0%	14	100%
4	***CEP + Telefone***	**3**	**3**	**100%**	0	0%
5	***Telefone + CEP***	**9**	**9**	**100%**	0	0%
8	***Endereço Básico + CEP + CidadeEstado***	**2**	**2**	**100%**	0	0%
10	*Endereço Básico + Telefone*	116	115	99,3%	1	0,9%
11	*Endereço Básico +Telefone + CidadeEstado*	21	3	14,3%	18	85,7%
12	*Endereço Básico + CidadeEstado + Telefone*	4	2	50%	2	50%
13	*Endereço Básico + CidadeEstado*	192	3	1,6%	189	98,4%
14	***Endereço Básico + CEP + Telefone***	**4**	**4**	**100%**	0	0%
15	***Endereço Básico + Telefone+ CEP***	**6**	**6**	**100%**	0	0%
16	***Endereço Básico + CEP***	**5**	**5**	**100%**	0	0%
17	*CEP*	26	4	15,3%	22	84,6%
	TOTAL	**549**	**280**	**4,34%**	**269**	**4,03%**

No caso do padrão *CEP* (17), o problema foi devido a uma das expressões inicialmente definidas para reconhecimento do CEP, que considera o código no formato "nn.nnn.nnn", sem que ele seja precedido da palavra CEP. Esse padrão extraiu outros números como CNPJ, endereço IP e população. Quando o CEP está no contexto do endereço, este tipo de erro não ocorre mas, procurar apenas pelo padrão *CEP* corre-se o risco de gerar inconsistências na geocodificação, caso exista para o número equivocado um CEP correspondente no *Locus*. Para evitar este problema, essa expressão foi

eliminada do padrão *CEP* e o extrator alterado. Após a alteração, executando novamente o extrator verificou-se que este tipo de problema foi evitado.

No caso dos padrões *CidadeEstado* + *Telefone* (2) e *Telefone* + *CidadeEstado* (3), o problema estava no nome da cidade extraído pois, normalmente, o número de telefone extraído estava correto. No caso do padrão *Endereço Básico* + *CidadeEstado* (13), o problema também estava no reconhecimento do nome da cidade.

Após a análise do conteúdo extraído, os principais erros no reconhecimento e extração de nomes de cidade podem ser assim resumidos: (1) falta de separador entre o nome do bairro e o nome da cidade e (2) cidades que têm o mesmo nome do estado onde se localizam (Rio de Janeiro e São Paulo). A seguir estão listados e exemplificados[45] os problemas detectados.

1 - O nome de um local de referência aparece no lugar do nome da cidade. No endereço "Rua .....- Posto Portuário de Ihéus – BA", "Posto Portuário de Ihéus" é reconhecido como o nome da cidade.

2 - A falta de separador entre o nome do bairro e o nome da cidade faz com que não seja possível individualizá-los. No endereço "Rua São Mateus 644 Lj-15; São Mateus Juiz de Fora – MG", "São Mateus Juiz de Fora" é reconhecido como nome da cidade.

3 - As cidades do Rio de Janeiro e São Paulo possuem o mesmo nome do estado, o que leva a dois problemas: (1) quando o endereço não inclui o nome do estado, só o nome da cidade, faz com que o nome da cidade seja interpretado como o nome do estado e o termo que o antecede como o nome da cidade; (2) quando o endereço inclui o nome do estado, ele não é reconhecido. O nome da cidade é reconhecido como o nome do estado e o termo que o antecede como o nome da cidade. No endereço "Rua da Reitoria 74, Cidade Universitária, São Paulo", "Cidade Universitária" é reconhecida como o nome da cidade e "São Paulo" como nome do estado. No endereço "Av. Nilo Peçanha, 50 – 3º andar- sala 3215 – Centro – Rio de Janeiro – RJ", "Centro" é reconhecido como nome da cidade e "Rio de Janeiro" como nome do estado.

4 - Presença do nome da região metropolitana no lugar do nome da cidade. No endereço "Rua Prefeito Reinaldo Alves 25 Grande Florianópolis – SC", "Grande Florianópolis" é reconhecida como o nome da cidade.

[45] Os endereços nos exemplos estão descritos conforme encontrados nas páginas da Web

5 - Abreviatura do nome de cidade com ponto. O ponto é interpretado como um separador fazendo com que o termo antes do ponto fique no complemento. No endereço "Sto. André / SP", "André" é reconhecido como o nome da cidade e Sto. como complemento.

Para minimizar os erros de reconhecimento listados acima, é necessário tratar de forma diferente os padrões referentes às cidades do Rio de Janeiro e São Paulo. Também deve ser considerada a existência de abreviaturas padronizadas como parte do nome de cidades (por exemplo, Sto., Sta. S.). Nos casos onde o nome da cidade vem antecedido de outro nome, como, por exemplo, Grande Florianópolis, poderá ser verificado no momento da geocodificação se, para o mesmo estado, existe uma cidade que corresponda a uma parte do nome reconhecido.

Observamos ainda que, nos padrões onde *CidadeEstado* aparece precedido pelo *CEP* ou por um número de telefone, existe uma possibilidade menor de que o conteúdo extraído esteja incorreto, pois não existe um texto para ser confundido com o nome da cidade. A presença de outros identificadores de localização, além de *CidadeEstado*, traz a vantagem de suprir uma informação errada. Por exemplo, se o nome da cidade estiver escrito de uma forma incorreta ou tiver sido reconhecido incorretamente, os outros identificadores podem prover a informação necessária.

Quanto aos números de telefone extraídos incorretamente, os motivos identificados foram dois: (1) o número de telefone é confundido com outro número, o que só acontece quando é usado o padrão *Telefone*; (2) o código de área é reconhecido erroneamente, como no exemplo "Rod. Br 422 s/n Km 74 3532186", onde "74" é interpretado como DDD.

As inconsistências encontradas entre os nomes de cidade e estado extraídos e aqueles obtidos através da geocodificação também foram avaliadas. Dentro da amostragem considerada, os problemas encontrados foram decorrentes de: (1) divergência entre o DDD informado para o telefone e a cidade considerada, quando a geocodificação foi feita pelo número de telefone; (2) nome da cidade com conteúdo errado (conforme já discutido) e a geocodificação feita usando-se o CEP ou o número de telefone.

5.2.5.2 Avaliação do Conteúdo - Geocodificação dos Endereços

Esta etapa da avaliação fez uso do resultado da seção anterior e analisou, para as ocorrências geocodificadas pelos identificadores de localização, a qualidade dos endereços extraídos. A avaliação do endereço foi feita por meio de sua geocodificação. Para essa geocodificação foram usadas duas abordagens: geocodificação automática e geocodificação manual. Para a geocodificação automática considerou-se a disponibilidade de um banco de dados geográfico com endereços e para a geocodificação manual a disponibilidade de mapas digitais com a presença de endereços no *software* MapLink[46].

A geocodificação automática foi feita para endereços na cidade de Belo Horizonte devido à disponibilidade do banco de dados geográfico da cidade com infra-estrutura de endereçamento. Usando os nomes de cidade e estado atribuídos pelo *Locus*, foram selecionadas todas as extrações de Belo Horizonte que continham o *Endereço Básico* (*Tipo do Logradouro, Nome do Logradouro* e *Número do Imóvel*) preenchido. Ao todo, foram encontrados, nas 385 páginas, 245 endereços extraídos por diferentes tipos de padrão. Exemplos desses endereços são mostrados na Figura 5.10. Nela podemos ver, entre outros, o tipo de padrão usado na extração, o tipo e nome do logradouro, o número do imóvel e o CEP extraídos, e o nome da cidade e a sigla do estado atribuídos pelo *Locus*.

Utilizando o banco de dados geográfico de Belo Horizonte, foram geocodificadas 233 ocorrências pelo endereço exato, 10 pelo endereço aproximado e 2 não foram geocodificadas. Em uma delas, o endereço era de Betim e o telefone de Belo Horizonte. Na outra, o endereço era de São Paulo e o telefone de Belo Horizonte. Apesar de, na geocodificação usando o número de telefone, as extrações terem sido identificadas como sendo de Belo Horizonte, na localização do endereço foi possível reparar o erro. Na Figura 5.10, pode ser visualizado parte do mapa de Belo Horizonte, com o resultado da geocodificação de endereços. Considerando a amostra de endereços de Belo Horizonte, podemos considerar que 99% dos endereços foram geocodificados corretamente, o que demonstra que a extração dos componentes do endereço foi extremamente eficaz.

[46] `http://maplink.uol.com.br`

Tipo_extracao	Tipo_logradouro	Nome_logradouro	Numero	Nome_cidade_locus	Sigla_estado_locus	Cep
6	Rua	dos Tamoios	00666	BELO HORIZONTE	MG	30,130,140
6	Rua	Santa Catarina	00941	BELO HORIZONTE	MG	30,170,080
6	Ave	Bernardo Monteiro	00918	BELO HORIZONTE	MG	30,150,281
6	Rua	Lhana	00075	BELO HORIZONTE	MG	30,410,460
6	Rua	Curitiba	00832	BELO HORIZONTE	MG	30,170,120
7	Ave	Brasil	00803	BELO HORIZONTE	MG	30,140,000
7	Ave	Olegario Maciel	00311	BELO HORIZONTE	MG	30,180,110
7	Ave	Olegario Maciel	00311	BELO HORIZONTE	MG	30,180,110
7	Ave	Olegario Maciel	00311	BELO HORIZONTE	MG	30,180,110
7	Rua	dos Goitacazes	00333	BELO HORIZONTE	MG	30,190,911
7	Ave	Brasil	00803	BELO HORIZONTE	MG	30,140,000
7	Ave	Ave do Contorno	09636	BELO HORIZONTE	MG	30,110,140
7	Rua	Rua Dos Aimores	02480	BELO HORIZONTE	MG	30,140,072
7	Ave	Ave Barbacena	01018	BELO HORIZONTE	MG	30,190,000
7	Ave	Francisco Sales	01017	BELO HORIZONTE	MG	30,150,221
7	Ave	do Contorno	04023	BELO HORIZONTE	MG	30,110,090
7	Rua	Domingos Vieira	00587	BELO HORIZONTE	MG	30,150,240
7	Ave	Olegario Maciel	00311	BELO HORIZONTE	MG	30,180,110
8	Ave	Ave Olegario Maciel	02360	BELO HORIZONTE	MG	30,180,112
8	Ave	Joao Pinheiro	00100	BELO HORIZONTE	MG	30,130,010
8	Rua	Curitiba	00832	BELO HORIZONTE	MG	30,170,010
8	Rua	Curitiba	00832	BELO HORIZONTE	MG	30,170,120
8	Rua	Paraiba	00697	BELO HORIZONTE	MG	31,130,140
8	Ave	Afonso Pena	01534	BELO HORIZONTE	MG	30,130,005
8	Ave	Alvares Cabral	00200	BELO HORIZONTE	MG	30,170,000
8	PCA	Hugo Werneck	00253	BELO HORIZONTE	MG	30,150,300
8	Ave	Afonso Pena	01212	BELO HORIZONTE	MG	30,130,908
8	Rua	Alvares Maciel	00059	BELO HORIZONTE	MG	30,150,250
8	Rua	Rua Claudio Manoel	01162	BELO HORIZONTE	MG	30,140,100
8	Rua	Espirito Santo	01059	BELO HORIZONTE	MG	30,160,922
8	Rua	dos Carijos	00528	BELO HORIZONTE	MG	30,120,060
8	Ave	Professor Alfredo Balena	00190	BELO HORIZONTE	MG	30,130,100
8	Rua	Rua Amethista	00025	BELO HORIZONTE	MG	30,410,420

Figura 5.10 – Área central de Belo Horizonte mostrando a localização dos endereços extraídos.

Considerando que para outras cidades não existia um banco de dados geográfico disponível, optou-se por executar a geocodificação de forma manual usando o serviço "localize um endereço" do MapLink. Assim, poderia-se dar mais credibilidade à

extração de endereços e mostrar que, por meio da integração com serviços Web como o MapLink, é possível geocodificar endereços de várias cidades do Brasil.

Para a geocodificação manual, foram selecionados da amostragem todos os endereços extraídos de algumas cidades que estão na área de cobertura do MapLink. As cidades escolhidas foram: Rio de Janeiro, Belém, Curitiba, São Paulo, Florianópolis, Salvador e Campinas, totalizando 847 endereços. O procedimento utilizado é descrito a seguir.

Cada endereço foi submetido ao MapLink para verificar a sua existência. Caso o endereço fosse localizado pelo MapLink, ele seria considerado correto e a geocodificação exata. Se o logradouro fosse encontrado no Maplink mas o número do imóvel não, seria pesquisado um número próximo. Se encontrado, o endereço seria considerado correto e a localização aproximada. Se o endereço não fosse encontrado, seria pesquisado se no endereço extraído existia um CEP. Se existisse, o CEP seria consultado no *site* dos Correios[47] para verificar a existência do logradouro. Se o CEP fosse válido e o logradouro existisse, o endereço e a extração seriam considerados corretos, porém ele não seria geocodificado devido à desatualização do banco de dados geográfico usado pelo MapLink. A Tabela 5.10 mostra o resultado da geocodificação de endereços e a Figura 5.11 ilustra a localização de um endereço de Curitiba no MapLink.

Tabela 5.10– Resultado da geocodificação de endereços

CIdades	Total de Endereços Extraídos	Endereços com Localização Exata	Endereços com Localização Aproximada	Endereços Validados mas não Localizados	Endereços não Localizados e não Validados
Belo Horizonte	245	233	10		2
Rio de Janeiro	145	134	3	4	4
São Paulo	333	320	4	2	7
Curitiba	293	283	1	4	5
Belém	10	10			
Florianópolis	9	8		1	
Salvador	34	31	1		2
Campinas	23	16	1		7
Total	**1092**	**1034 94,6%**	**20 1,8%**	**11 1%**	**27 2,5%**

[47] http://www.correios.gov.br

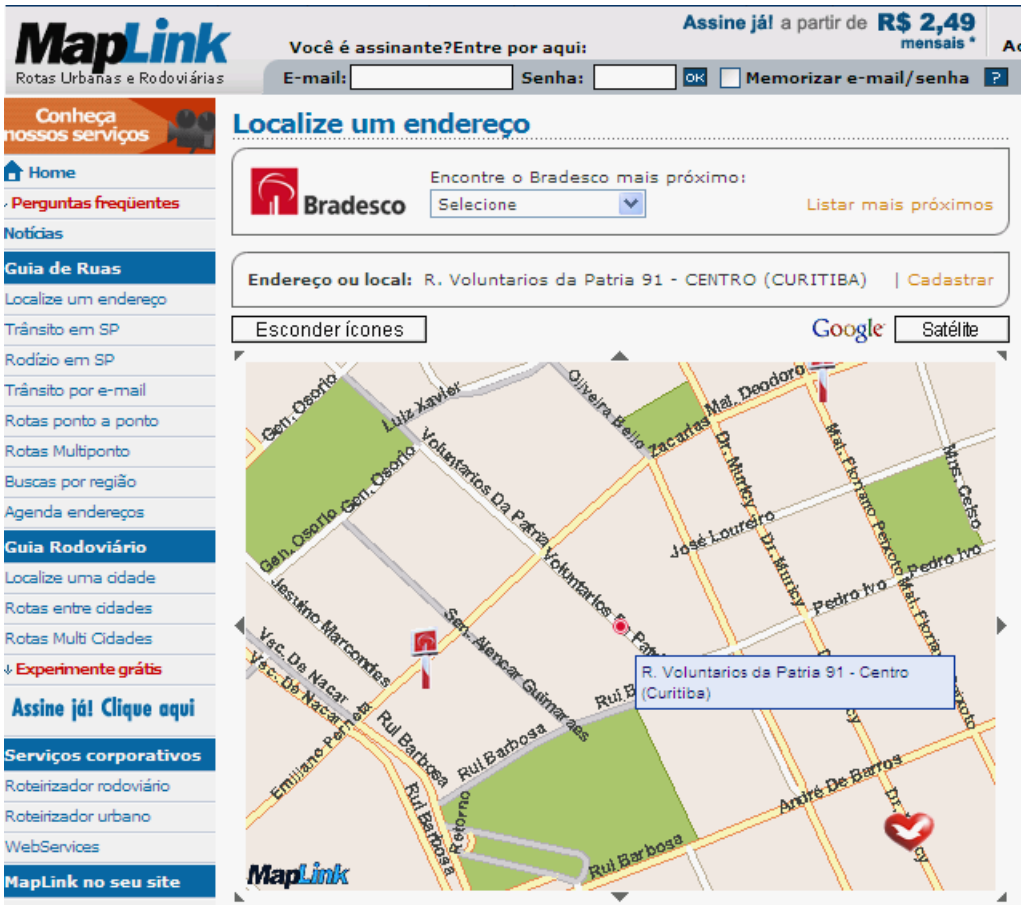


Figura 5.11 – Consulta a um endereço no MapLink

Como pode ser observado, o resultado da geocodificação automática e manual de endereços foi bastante satisfatório mostrando que os padrões definidos para extração de endereços são eficazes. Poucos endereços deixaram de ser localizados. Dos 1092 endereços considerados, 94,6% foram localizados de forma exata, 1,8% de forma aproximada, 1% não foi localizado porque não existia endereço correspondente no banco de dados geográfico consultado, mas teve o nome do logradouro validado no *site* dos Correios, e 2,5% dos endereços não puderam ser localizados e nem tiveram o nome do logradouro validado.

Os endereços não localizados foram extraídos corretamente mas não foram localizados devido aos seguintes problemas: (1) nome de logradouro inexistente não coincidindo com o CEP informado, (2) o endereço não existia no banco de dados geográfico consultado e nem possuía CEP, impossibilitando assim confirmar a existência do logradouro no *site* dos Correios, e (3) o número do imóvel não existia no banco de dados geográfico e o mais próximo encontrado era distante demais para ser

considerado. Observou-se, ainda, que a informação de bairro quando existente no complemento pode ser útil para tratamento de ambigüidade entre nomes de logradouro de uma mesma cidade. Por exemplo, em Belo Horizonte existem duas ruas São Paulo, uma localizada no centro da cidade e a outra no bairro Urca.

5.2.5.3 Avaliação da Eficácia do Extrator

Os resultados obtidos nas seções anteriores fizeram com que o extrator sofresse algumas alterações. O extrator foi novamente executado nas 385 páginas da amostragem gerando 9150 extrações, com todos os padrões encontrados. O resultado da extração foi geocodificado para a atribuição de cidade e estado a cada ocorrência.

Todas as 385 páginas foram inspecionadas visualmente, identificando todos os formatos de endereço que deveriam ser reconhecidos pelo extrator. Foi gerada, para cada página, uma lista com os tipos de padrão encontrados e sua respectiva quantidade. Esse resultado foi comparado com o que foi realmente extraído, verificando-se qual o percentual de endereços reconhecidos e o que aconteceu nos casos em que isto não foi possível. Os resultados sumarizados estão na Tabela 5.11. Em negrito estão os seis padrões selecionados (Seção 5.2.4).

Tabela 5.11 – Diferença entre a quantidade existente nas páginas e a quantidade extraída

ID	Tipo de Padrão	Qte Existente	Qte Extraída	Percentual Extraído
1	***Telefone***	**4312**	**3163**	**73,35%**
9	***Endereço Básico + CidadeEstado + CEP***	**774**	**452**	**58,40%**
10	***Endereço Básico + Telefone***	**825**	**565**	**68,48%**
13	***Endereço Básico + CidadeEstado***	**1173**	**867**	**73,91%**
16	***Endereço Básico + CEP***	**1438**	**1253**	**87,13%**
17	***CEP***	**2559**	**2538**	**99,18%**
2	*CidadeEstado + Telefone*	224	35	15,63%
3	*Telefone + CidadeEstado*	38	2	5,26%
4	*CEP + Telefone*	297	2	0,67%
5	*Telefone + CEP*	11	4	36,36%
6	*Endereço Básico + CEP + CidadeEstado + Telefone*	378	2	0,60%
7	*Endereço Básico + CidadeEstado + CEP + Telefone*	275	0	0%
8	*Endereço Básico + CEP + CidadeEstado*	767	226	29,47%
11	*Endereço Básico + Telefone + CidadeEstado*	13	8	61,54%
12	*Endereço Básico + CidadeEstado + Telefone*	177	29	16,38%
14	*Endereço Básico + CEP + Telefone*	316	2	0,63%
15	*Endereço Básico + Telefone + CEP*	9	2	22,22%

Avaliando os resultados, percebeu-se que os percentuais encontrados com os seis padrões selecionados foram bastante satisfatórios. O padrão *CEP* (17) foi o que obteve o melhor percentual (99,18%), seguido pelo padrão *Endereço Básico* + *CEP* (16) (87,13%). O padrão *Telefone* (1) obteve um percentual muito bom considerando a existência de números de fax e de vários números de telefone associados a um mesmo DDD presente nas páginas. Na contagem manual dos números de telefone existentes, tanto os vários números de telefone quanto os de fax foram considerados, já o extrator considera apenas o primeiro número encontrado. Por exemplo, para os telefones (31) 33775566 / 33786134 / 32217578, foram considerados três números de telefone na contagem manual, enquanto o padrão usado pelo extrator considera somente o primeiro. Os piores resultados foram obtidos com os padrões *Endereço Básico* + *CEP* + *Telefone (14)* e *CEP* + *Telefone (4)* e com os dois padrões mais completos, *Endereço Básico* + *CEP* + *CidadeEstado* + *Telefone* (6) e *Endereço Básico* + *CidadeEstado* + *CEP* + *Telefone* (7).

Podemos observar, nos experimentos anteriores, que os dois padrões mais completos (6 e 7), tiveram uma representatividade baixa (Tabela 5.3), mas uma geocodificação alta (Tabela 5.4). Na Tabela 5.11, esses padrões tiveram um percentual baixo de reconhecimento, o que poderia justificar a baixa representatividade encontrada na Tabela 5.3. No entanto, apesar da dificuldade de reconhecimento de padrões maiores, quando o extrator consegue reconhecê-los, a extração é feita corretamente.

Além dos problemas já comentados na Seção 5.2.5.1, referentes ao reconhecimento do nome da cidade, os principais problemas encontrados no reconhecimento de endereços estão listados a seguir. Conforme será visto, muitos deles impedem o reconhecimento do endereço na forma mais completa, sendo reconhecido apenas parte dele.

1 - Endereços constituídos pelo nome de um local

Quando o tipo e o nome do logradouro são substituídos pelo nome de um local de referência, não é possível extrair o endereço de forma completa. Apenas são extraídos os componentes *CEP*, *CidadeEstado* e *Telefone*, quando eles existirem. Abaixo são listados alguns endereços encontrados em páginas da Web que exemplificam o exposto.

- "Centro de Filosofia e Ciências Humanas – 10 andar – depto. de Historia, cidade universitária, CEP 5067901, Recife – PE Fone (0xx81) 32718950"
- "Quinta da Boa Vista, São Cristóvão Rio de Janeiro RJ, Cep 20940-000"
- "Ed. Niemeyer 6º. Andar"
- "Aeroporto de Uberlândia s/n"

2 - Endereços de Brasília e Palmas

O tipo de endereçamento usado em Brasília e Palmas não é reconhecido pelos padrões que foram definidos. Nesse tipo de endereçamento, a parte referente ao *Endereço Básico* não é reconhecida. Por este motivo, apenas são extraídos os componentes *CEP*, *CidadeEstado* e *Telefone*, quando eles existirem. O endereço "SHCN CL QD 407, Bloco E, loja 63, parte A Subsolo - Asa Norte 70855-550 - Brasília – DF Fone: (61) 273-0032", exemplifica esta situação, mostrando que o tipo de logradouro SHCN não segue os tipos de logradouro definidos. Nesse endereço seriam reconhecidos os seguintes padrões: *CEP*, *CidadeEstado* + *Telefone* e *Telefone*.

3 - Endereços em loteamentos

Normalmente, a descrição de um endereço em novos loteamentos é feita em função do nome do loteamento, da quadra e do lote, não possuindo tipo e nome de logradouro. Nestes casos, apenas a parte do endereço referente aos identificadores de localização é extraída. No endereço "Loteamento São Judas Tadeu, Quadra M lote 19 Bairro São Jose, 45204-040 Jequié, Bahia, tel. res. (73) 525-0500", seriam reconhecidos os seguintes padrões: *CEP*, *CidadeEstado* + *Telefone* e *Telefone*.

4 - Endereços com complemento muito longo

Quando o tamanho do complemento ultrapassa o limite estipulado (trinta caracteres), não é possível reconhecer o endereço completo. Assim, o endereço é dividido em duas partes, uma que antecede o complemento e outra que o precede. A parte do endereço que antecede o complemento não será extraída porque não existe um identificador de localização associado e, nesse caso, o padrão não a reconhece como um endereço. Na parte que precede o complemento serão extraídos os componentes *CEP*, *CidadeEstado* e *Telefone*, quando existirem. Observe o endereço "KONTIK -

FRANSTUR VIAGENS E TURISMO Avenida Tancredo Neves, 969 - 10 andar - sala 1004 Edf. Metropolitan Center - Caminho das Árvores CEP 41.820-020 - Salvador - Bahia Tel.: (0**71) 3483-8686". Devido ao tamanho do complemento, o CEP não é associado à Avenida Tancredo Neves. Dessa forma, são reconhecidos os padrões: *CEP*, *CidadeEstado* + *Telefone* e *Telefone*

5 - Endereços com variações incomuns ou com a presença do nome do país entre os seus componentes

A introdução do nome do país entre os componentes de um endereço ou variações incomuns na posição desses componentes fazem com que apenas parte do endereço seja extraído. Como exemplo, podemos citar a presença do nome do bairro depois do CEP ou o nome do país antes do número de telefone. No endereço "Av. Bandeirantes, 3900 - CEP 14040-901 - Bairro Monte Alegre - Ribeirão Preto - SP - Brasil - Tel. 55-0xx16-602.3670", o nome do bairro aparece entre os componentes CEP e CidadeEstado quebrando a seqüência de reconhecimento prevista nos padrões definidos. Nesse caso, serão reconhecidos os padrões: *Endereço Básico* + *CEP*, *CEP*, *CidadeEstado* e *Telefone*. Já no endereço "Rua dos Mundurucus, 4487 – Guama Belém-Pará-Brasil - CEP: 66.073-000 Fax: (091) 249-5365", o termo "Brasil" aparece entre os componentes CidadeEstado e CEP, quebrando a seqüência de reconhecimento prevista.

6 - Endereços com nome de cidade e sem nome de estado

Alguns endereços que aparecem em formatos como "*Endereço Básico*+ *Cidade* + *Telefone*" ou "*Endereço Básico* + *Bairro* + *Cidade* + *Telefone*" são reconhecidos parcialmente dependendo do número de posições ocupadas pelo complemento. Como não existe o nome do estado junto ao nome da cidade, o nome da cidade será considerado complemento devido à existência de um identificador de localização logo após. Nos dois formatos, o nome da cidade, com ou sem o nome do bairro, ficará como complemento se ocupar menos de 30 posições, caso contrário, só é extraído o *Telefone* porque a parte referente ao *Endereço Básico* não será seguida de um identificador de localização. Observe o exemplo "Rua Norma Stefani 90 sala 23 Barbacena Tel.: 32 33314489". Como o complemento mais o nome da cidade ocupam menos de 30 posições, será extraído o seguinte endereço: Rua Norma Stefani 90 (Endereço Básico), sala 23 Barbacena (Complemento) e 32 33314489 (Telefone).

7 - Endereços com nome do logradouro mas sem o tipo do logradouro

Endereços que apresentam o nome do logradouro sem ser antecedido pelo tipo do logradouro só são recuperados parcialmente. O componente *Endereço Básico* não é reconhecido devido à falta do tipo. Apenas são extraídos os componentes *CEP*, *CidadeEstado* e *Telefone*, quando existirem. No endereço "Epitácio Pessoa 172 Loja 16 A Santos - São Paulo (013) 3271-1096", "Epitácio Pessoa" não é reconhecido como nome de logradouro. Neste caso, são reconhecidos os seguintes padrões: *CidadeEstado* + *Telefone* e *Telefone*.

8 – Endereços com a presença de mais de um número de telefone no meio do endereço

A existência de mais de um número de telefone na seqüência dos componentes de um endereço impede o reconhecimento de um padrão mais completo. No endereço "R: Cel. Joaquim Jose de Oliveira, 199, Taquaral, Campinas-SP Tel: (19) 3254-4029 e 3253-7893 Cep: 13090-170", são reconhecidos quatro padrões de endereço: *Endereço Básico* + *CidadeEstado* + *Telefone*, *Telefone*, *Telefone* + *CEP* e *CEP*.

9 – Endereços com números de telefone em formato incomum ou identificador de telefone não definido

Números de telefone em formato não convencional, como "0/xx/DDD/número", não são reconhecidos pelos padrões definidos. Por exemplo, o número de telefone "0/xx/21/2277-7474" não seria reconhecido por nenhum dos padrões que incluem o componente *Telefone*.

Um identificador de telefone não definido impede o reconhecimento do padrão completo. No endereço "R. Paraiba, 466 - Centro Cep: 65900-000 Plantao: (98) 977-6538", o telefone seria reconhecido sozinho, pelo padrão *Telefone*. O termo "Plantao" não foi definido como identificador de telefone quebrando a seqüência esperada de *CEP* + *Telefone*.

10 – Endereços com a presença do padrão "*Endereço Básico* + *CEP* + *CidadeEstado* + *Telefone*"

O padrão definido considera apenas os componentes: *Endereço Básico* + *CEP* + *CidadeEstado*. Nesse caso, o endereço poderia ser reconhecido por esse padrão e pelos padrões *CidadeEstado* + *Telefone* e *Telefone*. O endereço: "Rua Santa Lucia, 291

Bairro Aclimação 35930-117 João Monlevade-MG Telefax: (0 xx 31) 852-2970" exemplifica este caso.

11 – Endereços inconsistentes

Algumas vezes, o endereço informado pertence a uma cidade e o telefone a outra. No exemplo "SESC Bauru Av. Aureliano Cardia, Bauru – SP (41) 235-1750", o endereço é de Bauru em São Paulo e o DDD é do estado do Paraná. Outras vezes existe inconsistência na localização do endereço, como no exemplo "Rua Milton Campos, 202 - Vila da Serra - Nova Lima BELO HORIZONTE – MG CEP: 30112-970 Fone: (031) 286-2039 Fax: (031) 286-1538". O endereço é de Nova Lima, mas também inclui o nome da cidade de Belo Horizonte. Como o termo "Belo Horizonte" está junto ao nome do estado, "Nova Lima" ficaria no complemento.

12 – Endereços que incluem muitos marcadores HTML entre os seus termos

No caso do extrator implementado nesta tese, os marcadores HTML foram substituídos pelo separador <S>. O reconhecimento e a extração de um endereço completo depende do número de separadores <S> existentes entre os seus componentes. Este número foi limitado a dez, tendo sido escolhido após várias tentativas. Um limite maior aumenta a probabilidade de erro, reconhecendo, por exemplo, componentes que não fazem parte do endereço considerado. Se o número de separadores ultrapassar a dez, o endereço será extraído em partes. Para exemplificar, observe as Figuras 5.12 e 5.13. A Figura 5.12 mostra um fragmento de uma página da Web onde aparece o endereço "Avenida Cula Mangabeira, 211 CEP: 39401-022 (38) 3229 3000 (38) 3221 9210" e a Figura 5.13 mostra este mesmo fragmento após a coletada e o pré-processamento da página correspondente, onde aparece o endereço da seguinte forma: "Avenida Cula Mangabeira, 211 <S> <S> <S><S><S><S> CEP: 39401-022<S> <S><S><S><S><S><S> <S><S> <S> <S> <S> <S> <S> <S><S>(38) 3229 3000<S><S> <S><S>(38) 3221 9210". Neste exemplo, serão reconhecidos um endereço com o padrão *Endereço Básico + CEP* e os telefones. A distância entre os componentes Endereço Básico e CEP é menor que dez, já a distância entre os componentes CEP e o Telefone é maior que dez, não sendo possível a associação entre os dois. A existência desse tipo de problema fez com que muitos padrões não fossem reconhecidos. Na contagem manual foi considerada a seqüência de aparecimentos dos componentes do endereço, independente do número de separadores <S>. Já o extrator

limitou a distância entre os componentes reconhecendo, assim, bem menos endereços. Por exemplo, no endereço da Figura 5.13, existe o padrão *CEP* + *Telefone*, mas a distância existente entre os seus componentes faz com que o extrator não considere a presença deste padrão.

			Rua Pereira Guimarães, 08 Centro		
37	Mateus Leme	Niceu Apolinário Lima	CEP: 35670-000	(31) 3535 1340	(31) 3535 1432
			Praça Bom Jesus, 99		
38	Matozinhos	Adão Pereira Santos	CEP: 35720-000	(31) 3712 1055	(31) 3712 1835
			Avenida Cula Mangabeira, 211		
39	Montes Claros	Jairo Ataíde Vieira	CEP: 39401-022	(38) 3229 3000	(38) 3221 9210
			Praça Bernardino de Lima, 80		
40	Nova Lima	Vitor Penido de Barros	CEP: 34000-000	(31) 3541 1323	(31) 3541 1326
			Praça Barão do Rio Branco, 12		

Figura 5.12 – Fragmento de uma página da Web
Fonte:http://www.comig.com.br

<S> <S> <S><S><S><S><S>CEP: 35720-000<S> <S><S><S><S><S><S> <S><S> <S> <S> <S> <S> <S><S>(31) 3712 1055<S><S> <S><S>(
3712 1835<S><S> <S> <S> <S><S>39<S><S> <S> <S><S><S>Montes Claros<S><S><S> <S> <S> <S><S>Jairo Ataide Vieira<S><S> <S>
<S><S>Avenida Cula Mangabeira, 211 <S> <S> <S><S><S><S><S> CEP: 39401-022<S> <S><S><S><S><S><S><S> <S><S> <S> <S> <S> <S> <S> <S
<S>(38) 3229 3000<S><S> <S><S>(38) 3221 9210<S><S> <S> <S> <S><S>40<S><S> <S> <S><S><S>Nova Lima<S> <S><S><S><S> <S> <S>
<S>Vitor Penido de Barros<S><S> <S> <S><S><S>Praca Bernardino de Lima, 80 <S> <S> <S><S><S>CEP: 34000-000 <S><S> <S> <
<S><S> <S> <S><S>(31) 3541 1323<S><S> <S><S>(31) 3541 1326<S><S> <S> <S> <S><S>41<S><S> <S> <S><S><S>Ouro Preto<S><S><
<S> <S> <S><S>Marisa Maria Xavier<S> <S><S> <S> <S><S><S>Praca Barao do Rio Branco, 12 <S> <S> <S><S><S><S> CEP:
35400-000<S> <S><S><S><S><S><S> <S><S> <S> <S> <S> <S> <S> <S><S>(31) 3559 3208<S><S> <S><S>(31) 3551 2901<S><S> <S> <

Figura 5.13 – Fragmento da página pré-processada

Alguns dos problemas identificados são inerentes ao ambiente Web onde não existem padrões e muitos erros de grafia acontecem. Outros podem ser evitados por meio de algumas medidas tais como: adaptação dos padrões definidos para atender endereços de Brasília e Palmas, identificação de endereços com nome de locais de referência, inclusão do nome do país como parte dos padrões e, principalmente, redução dos marcadores *HTML* para um único separador. Acredita-se que a redução dos marcadores *HTML* para um único aumentará muito a capacidade de reconhecimento do extrator. Muitos dos problemas identificados impediram o reconhecimento do endereço na forma mais completa, porém não impediram o reconhecimento de pelo menos parte do endereço. Esse resultado reforça o uso dos seis padrões selecionados que, por serem padrões menores, estão menos propensos a esses problemas.

5.3 Experimento 3 - Retrato da Web Brasileira com Base nos Seis Padrões Principais de Extração

Após a avaliação dos padrões e a constatação de que os seis padrões selecionados são capazes de extrair com eficácia e qualidade um maior número de endereços, o extrator com os seis padrões foi aplicado à coleção WBR05, composta de 4 milhões de páginas, coletadas de *sites* da Web brasileira, em março de 2005 [MPZC05]. Essa coleção reflete de forma significativa a Web brasileira.

Como resultado, 2.137.601 endereços foram localizados em 603.798 páginas, o que representa 14,77% do total de páginas, confirmando resultados anteriores em [Himm05, Mccu01]. Na WBR05, 12% das páginas possuem um ou mais números de telefone, 6.99% possuem um ou mais CEP e 9,53% possuem endereços. A geocodificação por meio do CEP foi a mais eficaz comprovando os resultados anteriores. A quantidade de números de telefone presentes na coleção é praticamente o dobro da de CEP, reforçando a idéia de que existem muitas páginas só com números de telefone, inclusive de fax. O padrão *Endereço Básico* + *CidadeEstado* apresentou o menor índice de geocodificação (77,07%), seguido pelo padrão *Endereço Básico* + *Telefone* (79,89%). Os motivos que impedem uma melhor geocodificação já foram discutidos nas seções anteriores. No padrão *Endereço Básico* + *CidadeEstado* + *CEP*, a geocodificação pelo *CEP* (99,33%) foi muito superior à obtida pelo identificador de localização *CidadeEstado* (68,20%), o que validou a estratégia de manter este padrão entre os seis selecionados. Devido aos problemas já discutidos, a divergência entre as geocodificações obtidas com *CEP* e com *CidadeEstado* ficou em 33,25%. Nesses casos, deve-se optar pela geocodificação obtida com o *CEP*. A Tabela 5.12 apresenta um resumo da extração, a Tabela 5.13 apresenta o total de páginas onde foram encontrados cada um dos seis padrões e a Tabela 5.14 detalha, para cada um dos seis padrões, a quantidade extraída e geocodificada de endereços.

Tabela 5.12 – Resumo da extração na WBR05

TOTAL DE PÁGINAS	4.088.692	
Total de Páginas com Endereços Reconhecidos	603.799	14,77%
Total de Endereços Reconhecidos	2.137.601	50,71% - *Telefone* (1) 1,63% - *Endereço Básico + CidadeEstado + CEP* (9) 4,65% - *Endereço Básico +Telefone* (10) 9,93% - *Endereço Básico + CidadeEstado* (13) 10,83% - *Endereço Básico +CEP* (16) 22,03% - *CEP* (17)

Tabela 5.13 – Total de páginas com a presença de cada tipo de padrão

ID	**TIPO DE PADRÃO**	**TOTAL DE PÁGINAS COM CADA PADRÃO**	
1	*Telefone*	505.189	12,0%
9	*Endereço Básico + CidadeEstado + CEP*	24.475	0,60%
10	*Endereço Básico + Telefone*	55.244	1,35%
13	*Endereço Básico + CidadeEstado*	5.063	3,79%
16	*Endereço Básico + CEP*	154.761	3,79%
17	*CEP*	285.999	6,99%

Tabela 5.14 – Total de endereços extraídos e geocodificados

ID	**Tipo de Padrão**	**Total Extraído**	***Telefone***	***CEP***	***CidadeEstado***	
			Total Geocodificado (% Geocod.)	Total Geo-codificado (% Geocod.)	Total Geo-codificado (%Geocod.)	% Geo-codificado diferente do CEP
1	*Telefone*	1.083.913	865.966 (79,89%)			
9	*Endereço Básico + CidadeEstado + CEP*	34.832		34.600 (99,33%)	23.757 (68,20%)	33,25%
10	*Endereço Básico + Telefone*	99.297	80.884 (81,46%)			
13	*Endereço Básico + CidadeEstado*	217.274			167.488 (77,07%)	
16	*Endereço Básico + CEP*	231.406		229.851 (99,33%)		
17	*CEP*	470.879		453.380 (96,28%)		
	Total	**2.137.601**				

A presença de um CEP sugere a existência de um endereço. Uma avaliação das páginas onde foram feitos reconhecimentos com o padrão *CEP* mas não com o padrão *Endereço Básico* + *CEP* deve ser realizada com o intuito de descobrir os motivos pelos quais os endereços não foram reconhecidos e, assim, diminuir a diferença entre a quantidade extraída com o padrão *CEP* (285.999) e com o padrão *Endereço Básico* + *CEP* (154.761). A adaptação dos padrões para reconhecimento de endereços das cidades de Brasília e Palmas também contribuirá para diminuir essa diferença.

5.4 Conclusões

Neste capítulo foram apresentados e discutidos os resultados dos experimentos que validaram as idéias propostas nesta tese, possibilitando conhecer a estrutura e a incidência de endereços em páginas da Web brasileira. Os experimentos comprovaram a viabilidade de se efetuar a extração automática e a geocodificação de endereços encontrados em páginas da Web. Os resultados obtidos foram bastante satisfatórios, mostrando que os seis padrões selecionados são suficientes para o reconhecimento de endereços. O uso de padrões teve como vantagem eliminar o trabalho de se examinar o texto no entorno de cada termo reconhecido para verificar se é ou não parte de um endereço. A combinação dos identificadores de localização com o *Endereço Básico* fez com que as extrações fossem mais precisas, reduzindo o número de falsos endereços.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Revisão do Trabalho

Esta tese propôs uma ontologia de lugar urbano, denominada *OnLocus*, para auxiliar o processo de reconhecimento e extração de evidências geo-espaciais existentes em páginas da Web que permitam a geocodificação automática dessas páginas. A ontologia, classificada como ontologia de extração, deu origem a uma base de conhecimento sobre lugares do Brasil necessária não só para a atribuição de coordenadas geográficas aos lugares extraídos das páginas da Web, como também para resolver ambigüidades e validar os locais encontrados nas páginas. Apesar de a ontologia *OnLocus* definir várias formas de reconhecimento de um lugar, esta tese explorou as evidências de características locais, como endereços, códigos postais e telefones encontrados em páginas da Web, que poderiam, por exemplo, ser utilizadas pelas máquinas de busca para recuperação de páginas referentes a serviços e atividades em uma determinada localidade ou próximos a ela. Além disso, como será visto na seção 6.2, outras aplicações poderiam se beneficiar dessa abordagem. O foco desta tese foi a chamada *Web local*, propondo uma abordagem, apoiada em uma ontologia de lugar urbano, que permitiu reconhecer, extrair e geocodificar evidências geo-espaciais sob a forma de endereços encontrados em páginas da Web. O endereço é a evidência geo-espacial mais adequada às aplicações de busca local por estar intimamente associado a um local urbano e representar a localização física de serviços e atividades presentes nas páginas da Web.

Após uma exaustiva avaliação visual de páginas da Web brasileira com presença de endereços, foram propostos dezoito padrões que refletem as diversas maneiras encontradas para se descrever um endereço. Os dezoito padrões, definidos por meio de expressões regulares, foram testados em uma coleção com 75.413 páginas da Web brasileira onde foram avaliadas a presença, incidência e qualidade da extração de cada

um deles. Como resultado, observou-se que apenas seis dos dezoito padrões eram suficientes para o processo de reconhecimento e extração de endereços.

Tendo se mostrado eficaz, o conjunto dos seis padrões selecionados foi, então, aplicado à coleção WBR05, contendo 4 milhões de páginas, possibilitando uma visão geral da presença de endereços na Web brasileira. O resultado mostrou que aproximadamente 15% das páginas continham alguma forma de endereço segundo os padrões considerados, sendo que 6,99% delas continham pelo menos um CEP, 12% números de telefone e 9,53% endereços na forma básica, comprovando que a Web é realmente uma rica fonte de conteúdo local e que há muitas vantagens em se usar um endereço ou componentes dele como chave de acesso a esse conteúdo, especialmente quando o conteúdo é relativo a serviços e atividades da vida diária.

Como resultado do processo de extração e reconhecimento de endereços, obteve-se um Repositório de Lugares contendo, entre outros, a URL da página, o endereço extraído separado por componentes, o nome da cidade, o nome do estado, as coordenadas geográficas do RME da cidade e, quando possível, as coordenadas geográficas do local associado ao endereço. Como visto, a geocodificação dos lugares reconhecidos é feita em dois níveis de granularidade: um municipal (retângulo mínimo envolvente do município) e outro local, com a determinação das coordenadas geográficas associadas aos endereços. Dessa forma, toda a extração pode ser geocodificada. Os endereços que não puderam ser localizados geograficamente possuem pelo menos a geocodificação do município. Esse repositório provê a informação necessária para o desenvolvimento de um índice espacial e de outros tipos de índice que utilizem a informação de cidade e estado obtida, trazendo vantagens em relação à indexação tradicional das máquinas de busca por usar vários tipos de evidência geo-espacial para o reconhecimento de uma cidade.

Assim, as principais contribuições desta tese foram (1) definição de uma ontologia de lugar urbano *OnLocus*, para auxílio no processo de reconhecimento e extração de evidências geo-espaciais em páginas da Web que permitissem a geocodificação automática dessas páginas; (2) criação de uma base de conhecimento de lugares (*gazetter*) do Brasil baseada na ontologia de lugar urbano e que foi utilizada para implementação de um sistema de localização espacial denominado *Locus* [SDBD05, SDBD04, Souz05]; (3) caracterização dos endereços em páginas da Web como fontes

de evidência geo-espacial [BLMS03, LBCM05] e definição de padrões para o seu reconhecimento e extração; (4) proposta de uma estratégia de categorização geográfica de uma página, ou de partes dela, dentro da divisão territorial (País/Estado/Cidade); e (5) avaliação das características qualitativas e quantitativas dos endereços presentes nas páginas da Web brasileira.

6.2 Trabalhos Futuros

O trabalho apresentado nesta tese abre inúmeras possibilidades para exploração do conteúdo geográfico que vão além da busca local. Os trabalhos propostos para a continuação desta tese possibilitarão que as máquinas de busca possam, por exemplo, utilizar as evidências geo-espaciais encontradas em páginas da Web para melhorar o resultado de uma busca, categorizar geograficamente as páginas, possibilitar a navegação a partir de *links* geográficos ou mesmo visualizar o resultado de consultas em serviços como *Google Earth* e *Virtual Earth*.

A Figura 6.1 exemplifica um ambiente de busca na Web onde o usuário submete sua consulta a uma máquina de busca e o resultado considera, quando necessário, a indexação espacial.

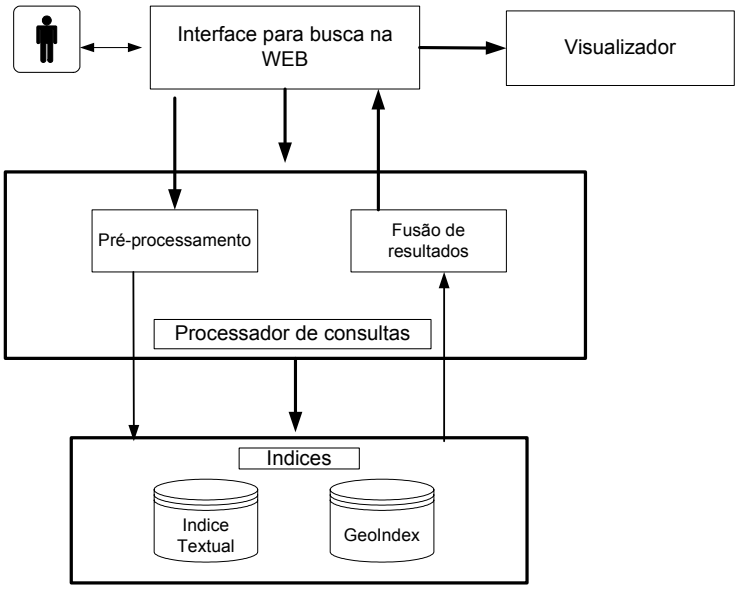


Figura 6.1 - Ambiente de busca na Web considerando o contexto geográfico

Se em uma consulta submetida à máquina de busca for verificada a existência de um nome de lugar em seu texto, ela passa por um pré-processamento para expansão desse nome de lugar. Termos relacionados ao mesmo lugar são também usados na busca. Por exemplo, a consulta "hotel 3 estrelas perto de Porto Seguro" poderia ter o nome "Porto Seguro" substituído pelas praias próximas a Porto Seguro. A busca é feita utilizando tanto o índice textual quanto o geográfico. Uma fusão do resultado obtido com o uso dos dois índices permite ordenar a consulta considerando a relevância geográfica. No exemplo da consulta a hotéis perto de Porto Seguro, as páginas de hotéis retornadas poderiam vir ordenadas pelo critério de proximidade em relação ao centro da cidade de Porto Seguro. De acordo com o tipo de consulta, a fusão pode considerar critérios de continência ou adjacência para a ordenação dos resultados. A apresentação do resultado de consultas com características geográficas considera a possibilidade de visualização em um mapa ou imagem de satélite. Na interface de consulta poderia também ser oferecida uma interface gráfica baseada em mapas ou imagens de satélite para consultas com contexto geográfico definido.

Portanto, sugerimos os seguintes trabalhos como continuação desta tese:

- Ampliação da capacidade de reconhecimento de endereços

A definição de novas heurísticas para ampliar o reconhecimento de endereços em outros formatos ou sem um identificador de localização aumentará a capacidade do extrator. Nesta tese não foram reconhecidos formatos de endereçamento como os de Brasília e os usados em loteamentos, que fogem ao padrão mais usual. Também não foram reconhecidos endereços constituídos pelo nome do lugar como, por exemplo, nomes de pontos de referência. Endereços sem um identificador de localização também não foram considerados e pelo que foi observado ainda é comum a presença de páginas onde os endereços não possuem um identificador de localização. Nesses casos, por exemplo, outras evidências poderiam ser utilizadas para o reconhecimento do local ao qual o endereço está associado. Como exemplo podemos citar os *links* e marcadores HTML especiais. Ao se analisar os *links* poderá ser verificado se páginas que apontam para essa página ou que são apontados por ela possuem algum contexto geográfico definido. No caso de uso de marcadores HTML especiais poderá ser verificado se no título existe algum nome que identifica o local ao qual a página se refere. Pontos de referência já cadastrados no *Locus* também poderiam ser usados para identificar de onde

é a página. Em páginas que possuem E*ndereço Básico* + *telefone* sem DDD, outras heurísticas podem ser desenvolvidas associando o prefixo do número de telefone e o nome do logradouro com os dados de Município e Logradouro existentes no *Locus*. É importante ressaltar que esse tipo de endereço ainda é muito comum em páginas de serviço existentes na Web.

- Integração de todas as formas de reconhecimento de lugar previstas na *OnLocus* em um único processo de reconhecimento e extração de evidências geo-espaciais

O processo de reconhecimento integrado possibilitaria a identificação de um lugar utilizando diversas evidências geo-espaciais encontradas. A combinação dessas evidências ampliaria a capacidade de reconhecimento de um lugar em páginas da Web e daria maior confiabilidade ao processo de reconhecimento. Para completar as formas de reconhecimento de lugar previstas na *OnLocus* é necessária a definição de heurísticas para reconhecimento de nomes de lugar.

- Atribuição de um escopo geográfico para um *site* com base nos lugares identificados em suas páginas

Avaliando os lugares encontrados nas páginas de um *site* é possível atribuir um lugar à página como um todo. Usando a hierarquia da divisão territorial, essa atribuição poderia, por exemplo, considerar a maior divisão que incluir as divisões territoriais encontradas.

- Identificação de páginas de serviços e categorização dos serviços associados aos endereços encontrados

Como já mencionado, o endereço é considerado a evidência geo-espacial mais adequada às aplicações de busca local por estar intimamente associado a um local urbano e representar a localização física de serviços e atividades presentes nas páginas da Web. No entanto, nem todos os endereços encontrados nas páginas são referentes a serviços podendo ser, por exemplo, um endereço residencial. Assim, a identificação de páginas referentes a serviços possibilitaria a identificação dos endereços relacionados a eles. Essa identificação pode ser feita associando a estrutura da página à quantidade e ao formato dos endereços encontrados. É desejável também que a categoria do serviço possa ser associada ao endereço encontrado. Termos que identificam componentes de

um endereço próximos a palavras-chave que denotam tipos de serviço como, hotel, apart hotel, restaurante, bar, bar e restaurante, buteco, oficina mecância, etc. poderiam ser usados na associação da categoria do serviço.

- Associação do nome do prestador do serviço ao endereço encontrado

A identificação, em páginas da Web, do nome do prestador do serviço junto ao endereço de sua localização e a categoria de serviço ao qual ele pertence possibilitaria que a busca local funcionasse como páginas amarelas na Web. Para a identificação do nome do prestador do serviço podem ser utilizadas heurísticas como, por exemplo, palavras-chave que denotam tipos de serviço (flat, oficina mecânica, hospital) seguidas, ou precedidas, de termos que possuem a primeira letra em maiúsculo e que estejam próximas a endereços reconhecidos.

- Realimentação do *gazetteer* do *Locus* com os pontos de referência encontrados em páginas Web

A deficiência da presença de nomes intra-urbanos atualmente existente nos *gazetters* faz com que o seu uso no reconhecimento de nomes de lugar, dentro de um contexto urbano em páginas da Web, fique bastante limitado. A identificação do local (cidade e estado) ou endereço (se existir) de um ponto de referência encontrado nas expressões de posicionamento pode ser usado para atualizar as referências urbanas de uma cidade em um *gazetteer* como o *Locus*. Para isso podem ser usadas heurísticas para a identificação de endereços encontrados próximos às expressões de posicionamento. A identificação da categoria à qual pertence o ponto de referência encontrado fará com que seu armazenamento no *gazetteer* possa ser categorizado. Alguns nomes de pontos de referência possuem palavras que denotam de onde ele é e a qual categoria pertencem como, por exemplo, Santa Casa de Misericórdia do Rio de Janeiro que identifica a categoria Hospital e a cidade do Rio de Janeiro. A incidência dos pontos de referência encontrados em páginas da Web é uma forma de identificar quais são os mais relevantes para uma comunidade local.

- Melhoria da geocodificação usando números de telefone

Como visto, o número de telefone tem uma presença marcante nas páginas da Web. Entretanto, observou-se que muitos números de telefone com DDD ainda se encontram

desatualizados, sem a inclusão do oitavo dígito, fazendo com que a geocodificação usando o "DDD + prefixo" não encontre correspondente no *Locus*. Uma forma de amenizar essa falta de atualização seria, por exemplo, testar se a inclusão do oitavo dígito nos números de telefone, encontrados em regiões que já possuem oito dígitos, melhoraria a geocodificação. Novas heurísticas podem ser desenvolvidas considerando cada um dos prefixos permitidos dentro da região considerada. Para a validação da geocodificação poderia ser verificado se o nome do logradouro encontrado no endereço existe na cidade identificada. Para isso poderia ser usada a relação de logradouros do *Locus*.

- Definição e implementação de um índice geográfico (GeoIndex), possibilitando a incorporação da dimensão geográfica aos resultados das buscas

A definição de mecanismos de indexação e busca apropriados proporcionaria a uma máquina de busca a recuperação de páginas da Web por meio da sua localização geográfica. A criação de um índice geográfico que contenha as coordenadas geográficas dos endereços encontrados na Web possibilita buscas locais identificando, a partir de um local de interesse, páginas que fazem referência a algum tipo de serviço e que estão localizadas dentro de um determinado raio. Outros índices que permitam busca com característica espacial também podem ser criados, como, por exemplo, índices que fazem uso do conhecimento topológico como adjacência e continência. Um índice desse tipo permitiria apontar para páginas que fazem referência a estados adjacentes ou páginas que fazem referência a lugares contidos em um mesmo estado ou em uma mesma cidade. Para isso, um novo método para cálculo de relevância considerando múltiplos índices (textuais e geográficos) precisa ser desenvolvido.

- Expansão de consulta com critérios geográficos utilizando a ontologia lugar

Nomes alternativos ou nomes relacionados a uma determinada localização podem ser usados para a recuperação de páginas fazendo com que a busca não fique limitada a termos que formam o nome de um lugar. Por exemplo, páginas que citam "Lagoa da Pampulha", "Mineirão" e "Igreja da Pampulha" estão associadas a Belo Horizonte. A base de conhecimento de lugares do Brasil pode ser usada para a obtenção dessas associações, assim como para filtrar uma consulta com critérios geográficos, por exemplo, evitando ambigüidades na busca. Se o nome de uma cidade, identificado no

texto da consulta, existir em vários estados pode ser dada uma opção de escolha de um estado entre todos os estados onde aquele nome de cidade existir.

- Desenvolvimento de novas formas de navegação a partir de *links* baseados em localização geográfica

Usando o raciocínio espacial e a ontologia de lugar, poderiam ser associados às páginas *links* geográficos de proximidade com outras páginas, como, por exemplo, páginas referentes ao mesmo estado, a estados vizinhos, à mesma cidade e a cidades limítrofes, ou páginas que possuem endereços próximos de um endereço existente em uma determinada página. Com a categorização geográfica das páginas e a categorização de serviços nas páginas, poder-se-ia associar *links* de proximidade geográfica que tratam do mesmo assunto ou de assuntos correlacionados.

- Integração com *sites* de publicação de mapas como MapLink ou com ferramentas como o *Google Earth* e *Virtual Earth*

A visualização dos resultados de uma consulta com características geográficas é muito importante para a compreensão do seu contexto por parte do usuário. Assim, as coordenadas dos endereços reconhecidos poderiam ser usadas para que os resultados de uma consulta fossem visualizados por meio de ferramentas disponíveis na Web como, o *Google Earth*, *Virtual Earth* ou MapLink.

- Desenvolvimento ou personalização de uma ferramenta para descrição de ontologias que incorpore semântica geográfica

Ontologias com características espaciais possuem particularidades semânticas que diferem da descrição de uma ontologia convencional. Um exemplo é a propagação dos relacionamentos espaciais. Subconceitos herdam relacionamentos espaciais definidos para um conceito. Para exemplificar, observe os conceitos *Cidade* e *Estado*. Apesar de existir entre eles um relacionamento espacial "parte-de", o relacionamento de adjacência entre estados não deve ser herdado por *Cidade*. O fato de Minas Gerais ser adjacente ao Rio de Janeiro não garante que Belo Horizonte, que é "parte-de" Minas Gerais, seja adjacente ao Rio de Janeiro ou a algum outro estado. O desenvolvimento de uma ferramenta que reflita mais adequadamente essas particularidades geográficas permitiria

a definição de regras para restrição de propagação de alguns relacionamentos espaciais e para a derivação de relacionamentos espaciais.

- Anotação automática em páginas da Web

A Web Semântica é uma extensão da Web atual que visa adicionar semântica ao conteúdo de suas páginas criando um ambiente onde agentes de *software* e usuários possam trabalhar de forma cooperativa [BeHL01]. Para isso, a Web Semântica prevê que sejam feitas anotações nas páginas da Web com o objetivo de facilitar o processamento semântico da informação. Nesta tese, o uso da ontologia de extração *OnLocus* permitiu que fossem atribuídos rótulos espaciais a determinados termos reconhecidos em páginas da Web, gerando assim uma informação semântica sobreposta aos termos encontrados. Dessa forma, o desenvolvimento de uma ferramenta para anotação automática desse tipo de informação nas páginas da Web permitiria, dentro do contexto da Web Semântica, a exploração da semântica espacial implícita, possibilitando que a informação geo-espacial identificada possa ser utilizada por outros serviços Web que considerem o contexto geográfico.

Referências bibliográficas

ACFG01 Arprez, J. C., Corcho, O., Fernandez-Lopez, M., Gomez-Perez, A. WebODE: A Scalable Workbench for Ontological Engineering. In *Proceedings of the First International Conference on Knowledge Capture*, 21-23, Victoria, B.C., Canada, 2001. Disponível em <http://mkbeem.elibel.tm.fr/paper/KCAP2001-35.pdf>. Acesso em maio 2005.

Adel98 Adelberg, B. NoDoSE: A Tool for Semi-Automatically Extracting Structured and Semi-Structured Data from Text Documents. *SIGMOD Record*, 27 (2): 283-294, 1998.

ADLG04 ADL - Alexandria Digital Library Gazetteer, 2004. Disponível em <http://www.alexandria.ucsb.edu>. Acesso em: maio 2005.

AHSS04 Amitay, E., Har'EI, N., Sivan, R. Soffer. A.Web-a-Where: Geotagging Web Content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 273-280, Sheffield, UK, 2004.

ArMe98 Arocena, G. O., Mendelzon, A. O. WEBOQL: Restructuring Documents, Databases, and Webs. In *Proceedings of the 14th IEEE International Conference on Data Engineering*, 24-33, Orlando, Florida, USA, 1998.

ArSO00 Arikawa, M., Sagara, T., Okamura, K. Spatial Media Fusion Project. In *Proceedings of Kyoto International Conference on Digital Library: Research and Practice*, IEEE Computer Society Press, 75-82, Kyoto, Japan, 2000.

Barb99 Barbetta, P. A. *Estatística Aplicada às Ciências Sociais*, Editora da UFSC, 3ª.ed., Florianópolis, 1999.

BaFG01 Baumgartner, R., Flesca, S., Gottlob, G. Visual Web Information Extraction with Lixto.In n *Proceedings of the 26th International Conference on Very Large Database Systems*, 119-128, Rome, Italy, 2001.

BCGG99 Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., Shivakumar, N. Exploiting Geographical Location Information of Web Pages. In *Proceedings of The ACM SIGMOD Workshop on Web and Databases*, 91-96, Philadelphia, Pennsylvania, USA, 1999.

BEHH02 Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S., Stojanovic, L., Stojanovic, N., Studer, R., Stumme, G., Sure, Y., Tane, J., Volz, R., Zacharias, V. KAON - Towards a Large Scale Semantic Web. In *Proceedings of E-Commerce and Web Technologies, Third International Conference,* 304-313, Aix-en-Provence, France*,* 2002.

BeHL01 Berners-Lee, J., Hendler, J., Lassila, O. The Semantic Web. *Scientific American*, 184 (5): 34-43, 2001.

BHGS01 Bechhofer S., Horrocks, I., Goble, C., Stevens, R. OilEd: a Reason-able Ontology Editor for the Semantic Web**.** In *Proceedings of KI 2001: Advances in Artificial Intelligence, Joint German/Austrian Conference on Artificial Intelligence*, Vol. 2174 of LNCS, 396-408, Springer-Verlag, Vienna, Austria, 2001.

BLMS03 Borges, K. A. V, Laender, A. H. F., Medeiros, C. B., da Silva, A. S., Davis Jr., C. A. The Web as a Data Source for Spatial Database. In *Proceedings of the V Brazilian Symposium on Geoinformatics*. Campos do Jordão, SP, Brasil, 2003, in CD-ROM. Disponível em <http://www.geoinfo.info>. Acesso em fevereiro de 2004

BoDL01 Borges, K. A. V., Davis Jr., C. A., Laender, A. H. F. OMT-G: An Object-Oriented Data Model for Geographic Applications. *GeoInformatica*, 5(3): 221-260, 2001.

BYRN99 Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley, 1999.

CaMo99 Callif, M. E., Mooney, R. J. Relational Learning of Pattern-Match Rules for Information Extraction. In *Proceedings of the 16th National Conference on Artificial Intelligence and 11th Conference on Innovative Applications of Artifical Intelligence,* 328-334, Orlando, Florida, USA,1999.

Camp93 Campbell, J. The Role of Physical Objects in Spatial Thinking. In Eilan, N., McCarthy, R., Brewer, M. W. (eds.), *Problems in the Philosophy and Psychology of Spatial Representation*, Oxford: Blackwell 1993; reprinted by Oxford University Press, 65-95, 1999.

CaVa96 Casati, R., Varzi, A. C. The Structure of Spatial Localization. *Philosophical Studies*, 82 (2):205-239, 1996.

CDLP04 Cox, S., Daisey, P., Lake, R., Portele, C., Whiteside, A. (ed.). *OpenGIS® Geography Markup Language (GML) Implementation Specification.* Open Geospatial Consortium, Inc., 2004. Disponível em <http://www.opengeospatial.org/specs/>. Acesso em março de 2005

Chin97 Chinchor, N. *MUC-7 Named Entity Task Definition*, Version 3.5, 1997. Disponível em <http://www-nlpir.nist.gov/related_projects/muc/ proceedings/muc_7_proceedings/overview.html>. Acesso em outubro de 2005.

ChSM05 Chaves, M. S., Silva, M. J., Martins, B. A Geographic Knowledge Base for Semantic Web Applications. In *Proceedings of the 20th Braziliam Simpósio Brasileiro de Banco de Dados,* 40-54, Uberlândia, MG, Brasil, 2005.

ClFO93 Clementini, E., Di Felice, P., Oosterom, P. V. A Small Set of Formal Topological Relationships Suitable for End-User Interaction. In *Proceedings of Third International Symposium on Advances in Spatial Databases,* 277-295, Singapore, 1993.

CoFG03 Corcho, O., Fernández-López, M., Gómez-Pérez, A. Methodologies, Tools and Languages for Building Ontologies: Where is their Meeting Point? *Data Knowledge Engineering,* 46 (1): 41 - 64, 2003.

CrMe98 Crescenzi, V., Mecca, G. Grammars have Exceptions. *Information Systems* 23 (8): 539 – 565, 1998.

CrMM01 Crescenzi, V., Mecca, G., Merialdo, P. RoadRunner: Towards Automatic Data Extractiom from Large Web Sites. In *Proceedings of the 26th International Conference on Very Large Database Systems*, 109-118, Rome, Italy, 2001.

DaFB03 Davis Jr., C. A, Fonseca, F. T., Borges, K. A. V. A Flexible Addressing System for Approximate Geocoding. In *Proceedings of the V Brazilian Symposium on Geoinformatics*. Campos do Jordão, SP, Brasil, 2003, in CD-ROM. Disponível em <http://www.geoinfo.info>. Acesso em fevereiro de 2004

DeBL05 Delboni, T. M., Borges K. A.V., Laender A. H. F. Geographic Web Search based on Positioning Expressions. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval held in conjunction with ACM CIKM 2005*, 61-64, Bremen, Germany, 2005.

DEFS99 Decker S., Erdmann, M., Fensel, D., Studer, R. *Ontobroker: Ontology Based Access to Distributed and Semi-Structured Information*. In R. Meersman et al. (Eds.), Database Semantics: Semantic Issues in Multimedia Systems, 351-369, Kluwer Academic Publisher, Boston, MA, 1999.

Delb05 Delboni, T. M. *Expressões de Posicionamento como Fonte de Contexto Geográfico na Web*. Dissertação de Mestrado, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 2005.

DiGS00 Ding, J., Gravano, L., Shivakumar, N. Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th International Conference on Very Large Databases*, 545-556, Cairo, Egypt, 2000.

DoMC99 Domingue, J., Motta, E., Corcho, O. *Knowledge Modelling in WebOnto and OCML: A User Guide*, 1999. Disponível em <http://kmi.open.ac.uk/projects/webonto/user_guide.2.4.pdf>. Acesso em março 2004.

EBNF96 ISO/IEC14977:1996. *Information technology - Syntactic metalanguage - Extended BNF*. Disponível em <http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm>. Acesso em dezembro de 2005.

ECJL99 Embley, D. W., Campbell, D. M., Jiang, Y. S., Liddle, S. W., Lonsdale, D. W., NG, Y. -K, Smith, R. D. Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages. *Data Knowledge Engineering*, 31 (3): 227-251, 1999.

Egen02 Egenhofer, M. J. Toward the Semantic Geospatial Web. In *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems*, 1-4, McLean, Virginia, USA, 2002.

EgSh98 Egenhofer, M. J., Shariff, A. R. B. M. Metric Details for Natural-Language Spatial Relations. *ACM Transactions on Information Systems* 16 (4): 295-321, 1998.

Embl04 Embley, D. W. Toward Semantic Understanding – An Approach Based on Information Extraction Ontologies. In *Proceedings of the Fifteenth Australasian Database Conference*, Dunedin, New Zeland, 2004. Disponível em <http://citeseer.ist.psu.edu/632008.html>. Acesso em Março de 2005.

FEAC02 Fonseca, F. T., Egenhofer, M. J., Agouris, P., Câmara, G. Using Ontologies for Integrated Geographic Information Systems. *Transactions on GIS*, 6 (3): 231-257, 2002.

FHHM01 Fensel, D., Van Harmelen, F., Horrocks, I., McGuinness, D., Patel-Schneider, P. OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2): 38-44, 2001.

Fran96 Frank, A. U. Qualitative Spatial Reasoning: Cardinal Direction as an Example. *International Journal of Geographical Information Systems,* 10 (3): 269-290, 1996.

Free75 Freeman, J. The Modelling of Spatial Relations. *Computer Graphics and Image Processing*, 4 (2):156-171, 1975.

Frei00 Freitag, D. Machine Learning for Information Extraction in Informal Domains. *Machine Learning*, 39 (2-3): 169-201, 2000.

FuAJ03 Fu, G., Abdelmoty, A., Jones, C. B. Maintaining Ontologies for Geographical Information Retrieval on the Web. In *Proceedings of OTM Confederated International Conferences CoopIS, DOA, and OOBASE*, 934-951, Sicily, Italy, 2003. Disponível em <http://www.geo-spirit.org/reports.html#2003>. Acesso em março de 2004.

FuJA05 Fu, G., Jones, C. B., Abdelmoty, A. Building a Geographical Ontology for Intelligent Spatial Search on the Web. In *Proceedings of IASTED International Conference on Databases and Applications,* 167-172, Innsbruck, Austria, 2005. Disponível em <http://www.geo-spirit.org/reports.html#2005>. Acesso em janeiro de 2006.

GeFi92 Genesereth, M. R., Fikes, R. E. *Knowledge Interchange Format, version 3.0 Reference Manual*, Logic Group Report Logic-92-1 Computer Science Department, Stanford University, Stanford, California, USA, 1992. Disponível em <http://www.cs.umbc.edu/kse/kif/>. Acesso em março de 2004.

GeNS02 GEOnet Names Server. Disponível em <http://earth-info.nga.mil/gns/html>. Acesso em dezembro de 2005.

GNIS06 GNIS- Geographic Names Information System. Disponível em <http://geonames.usgs.gov/domestic/index.html>. Acesso em dezembro de 2005.

GrHL03 Gravano, L., Hatzivassiloglou, V. Lichtenstein, R. Categorizing Web Queries According to Geographical Locality. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management,* 325-333, New Orleans, LA, USA, 2003.

Grub92 Gruber, T. R. A Translation Approach to Portable Ontology Specifications. Technical Report KLS 92-7, Knowledge Systems Laboratory, Stanford University, Stanford, CA, 1992

Grub92a Gruber, T.R. ONTOLINGUA: A Mechanism to Support Portable Ontologies. Technical Report KSL91-66, Knowledge Systems Laboratory, Stanford University, Stanford, CA, USA, 1992. Disponível em <http://www.ksl.stanford.edu/software/ontolingua/>. Acesso em março de 2004.

Grub95 Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human and Computer Studies*, 43 (5-6): 907-928, 1995.

GTGN00 Getty Thesaurus of Geographic Names. Disponível em <http:// www.getty.edu/research/conducting_research/vocabularies/tgn/>.Acesso em dezembro de 2005.

Guti94 Güting, R. H. An Introduction to Spatial Database Systems. The *VLDB Journal,* 3 (4): 357-399, 1994.

HaMG97 Hammer, J., McHugh, J., Garcia-Molina, H. Semistructured Data: The TSIMMIS Experience. In *Proceedings of the First East-European Symposium on Advances in Databases on Information Systems*, 1-8, St. Petersburg, Russia, 1997.

HaMo03 Haarslev, V., Möller, R. Racer: A Core Inference Engine for the Semantic Web. In *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools, located at the 2nd International Semantic Web Conference,* 27–36, Sanibel Island, Florida, USA, 2003.

HeHL99 Heflin, J., Hendler, J., Luke, S. SHOE: A Knowledge Representation Language for Internet Application. *Technical Report CS-TR-4078 (UMIACS TR-99-71),* 1999. Disponível em <http://www.cs.umd.edu/ projects/plus/SHOE/pubs/#tr99>. Acesso em março 2004.

HeKS03 Heinzle, F., Kopczynski, M., Sester, M. Spatial Data Interpretation for the Intelligent access to Spatial Information in the Internet. In *Proceedings of the 21st International Cartographic Conference,* Durban, South Africa, 2003. Disponível em <http://www.geo-spirit.org/ reports.html#2003>. Acesso em março de 2004.

HiIs01 Hiramatsu, K., Ishida, T. An Augmented Web Space for Digital Cities. In *Proceedings of IEEE Symposium on Applications and the Internet*, 105-112, San Diego, CA, USA, 2001.

Hill00 Hill, L. L. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, 280-290, Lisboa, Portugal, 2000.

Himm05 Himmelstein, H. Local Search: The Internet is the Yellow Pages. *IEEE Computer*, 38 (2): 26-35, 2005.

HoHa00 Horrocks, I., van Harmelen, F. Reference Description of the DAML+OIL Ontology Markup Language, 2000. Disponível em <http://www.daml.org/2000/12/reference.html>. Acesso em março de 2004

HoPH03 Horrocks, I., Patel-Schneider, P. F., van Harmelen, F. From SHIQ and RDF to OWL: The Making of a Web Ontology Language. *Journal of Web Semantics*, 1 (1): 7-26, 2003. Disponível em <http://www.cs.vu.nl/~frankh/abstracts/JWS03.html>. Acesso em março de 2004.

HsDu98 Hsu, C.-N., Dung, M.-T. Generating Finite-State Transducers for Semi-Structured Data Extraction from The Web. *Information Systems*, 23 (8): 521-538, 1998.

JPRS02 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. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 387-388, ACM Press, Tampere, Finland, 2002. Disponível em <http://www.geo-spirit.org/reports.html #2002>. Acesso em fevereiro de 2003.

KaCL01 Kambayashi, Y., Cheng, K., Lee, R. (2001). Database Approach for Improving Web Efficiency and Enhancing Geographic Information Systems. In *Proceedings of the International Conference on Internet Information Retrieva l*, 159-176, Seoul, Korea, 2001.

KaCT99 Karp, R., Chaudhri, V., Thomere, J. XOL: An XML-Based Ontology Exchange Language, 1999. Disponível em <http://www.ai.sri.com/pkarp/xol> . Acesso em março de 2004

Khan00 Khan, L. R. *Ontology-based Information Selection*. PhD thesis, Faculty of the Graduate School, University of Southern Califórnia, 2000. Disponível em <http://www.utdallas.edu/research/esc/publications/lkhan_def.pdf>. Acesso em julho de 2005.

KiSD01 Kiryakov, A., Simov, K. I., Dimitrov, M. OntoMap: Portal for Upper-level Ontologies. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, *47-58,* Ogunquit, Maine, USA, 2001.

Kush00 Kushmerick, N. Wrapper Induction: Efficiency and Expressiveness. *Artificial Intelligence Journal*, 118 (1-2): 15-68, 2000.

LaRS02 Laender, A. H. F., Ribeiro-Neto, B., da Silva, A. S. DEByE - Data Extraction by Example. *Data and Knowledge Engineering*, 40 (2): 121-154, 2002.

Lars96 Larson, R. R. Geographic Information Retrieval and Spatial Browsing. Smith, L. C. and Gluck M. (Eds), *Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information*, 81-123, University of Illinois, Urbana, IL, USA, 1996. Disponível em <http://sherlock.berkeley.edu/geo_ir/part1.html>. Acesso em março 2005.

LaSw99 Lassila, O., Swick, R. R. Resource Description Framework (RDF) Model and Syntax Specification. Technical Report Disponível em <http://www.w3.org/TR/PR-rdf-syntax/>. Acesso em outubro 2005.

LBCM05 Laender, A. H. F., Borges, K. A. V., Carvalho, J. C. P., Medeiros, C. B., da 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***,** 23-48, IRM Press, Hershey, PA, USA, 2005.

LiPH00 Liu, L., Pu, C., Han, W. XWRAP: An XML-enable Wrapper Construction System for Web Information Sources. In *Proceedings of the 16th IEEE International Conference on Data Engineering,* 611-621, San Diego, Califórnia, USA, 2000.

LRST02 Laender, A. H. F., Ribeiro-Neto, B., da Silva, A. S., Teixeira, S. J. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record*, 31 (2): 84-93, 2002.

MaEg95 Mark, D. M., Egenhofer, M. J. Topology of Prototypical Spatial Relations Between Lines and Regions in English and Spanish. In *Proceedings* of Auto Carto 12, 245-254, Charlotte, North Carolina, USA, 1995.

MAHM03 Morimoto, Y., Aono, M., Houle, M. E., McCurley, K. S. Extracting Spatial Knowledge from the Web. In *Proceedings of the Symposium on Applications and the Internet*, 326-333, Orlando, FL, USA, 2003. Disponível em <http://mccurley.org/papers/SAINT03.pdf>. Acesso em março de 2004.

Mccu01 McCurley, K. S. Geospatial Mapping and Navigation of the Web. In *Proceedings of the Tenth International World Wide Web Conference*, 221-229, Hong Kong, China, 2001.

McHa04 McGuinness D. L., van Harmelen, F. OWL Web Ontology Language Overview. Disponível em <http://www.w3.org/TR/owl-features/>. Acesso em março de 2004.

MGGF03 Montello, D. R., Goodchild, M. F., Gottsegen, J., Fohl, P. Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition and Computation*, 3 (2,3): 185-204, 2003.

MiMG99 Mikheev, A., Moens, M., Grover, C. Named Entity Recognition Without Gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*,1-8, Bergen, Norway, 1999.

MKPB03 Manov, D., Kiryakov, A., Popov, B., Bontcheva, K., Maynard, D., Cunningham, H. Experiments with Knowledge for Extraction. In *Proceedings of the Human Language Technology Conference Workshop on Analysis of Geographic*, 1-9, Edmonton, Canadá, 2003. Disponível em <http://tangra.si.umich.edu/clair/HLT-NAACL03/georef/pdf/ws903.pdf>. Acesso em abril de 2005.

MPZC05 Modesto, M., Pereira Jr., Á., Ziviani, N., Castillo, C., Baeza-Yates, R. Um novo retrato da Web brasileira. In *Proceedings of XXXII Seminário Integrado de Software e Hardware*, 2005-2016, São Leopoldo, RS, Brasil, 2005.

MuMK01 Muslea, I., Minton, S., Knoblock, C. A. Hierarchical Wrapper Induction for Semistructured Information Sources. A*utonomous Agents and Multi-Agent Systems*, 4 (1-2): 93-114 , 2001.

NiPe01 Niles, I., Pease, A. Towards a Standard Upper Ontology. In *Proceedings 2nd International Conference on Formal Ontology in Inforamtion Systems,* 2-9, Ogunquit, Maine, USA, 2001.

NSDC01 Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., Musen, M. A. Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16 (2): 60-71, 2001. Disponível em <http://hcs.science.uvanl/Capita-AI/2002/papers/Noy.pdf>. Acesso em maio 2003.

PuEg88 Pullar, D. V., Egenhofer, M. J. Towards the Defaction and Use of Topological Relations Among Spatial Objects. In *Proceedings of the 3rd International Symposium on Spatial Data Handling,* 225-242, Columbus, Ohio, USA, 1988.

RaBB03 Rauch, E., Bukatin, M., Baker, K. A Confidence-based Framework for Disambiguating Geographic Terms. In *Proceedings of the Human Language Technology Conference Workshop on Analysis of Geographic*, 50-54, Edmonton, Canadá, 2003. Disponível em <http://people.mokk.bme.hu/~kornai/NAACL/WS9/ws917.pdf>. Acesso em março de 2004.

Roch03 Roche C. Ontology: Survey. In *Proceedings of the* 8th *Symposium on Automated Systems Based on Human Skill and Knowledge*, 22-24, Goteborg, Sweden, 2003. Disponível em <http://ontology.univ-savoie.fr/condillac/en/ressources/download/documents/IFAC%202003.pdf>. Acesso em maio de 2005.

Rodr02 Rodríguez, M. A. T. A Spatial Dimension for Searching the World Wide Web. In *Proceedings of the Hybrid Intelligent Systems 2002*, 583-592, Santiago, Chile, 2002.

SaAz00 Sahuguet, A., Azavant, F. Building Intelligent Web Applications Using Lightweight Wrappers. *Data and Knowledge Engineering*, 36 (3), 283-316, 2001.

SaKo04 Sanderson, M., Kohler, J. Analyzing Geographic Queries. In *Proceedings of the ACM SIGIR Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004. Disponível em <http://www.geo.unizh.ch/~rsp/gir/abstracts/sanderson.pdf>. Acesso em maio de 2005.

SCPV04 Spaccapietra, S., Cullot, N., Parent, C., Vangenot, C. On Spatial Ontologies, *VI Brazilian Symposium on Geoinformatics*. Campos do Jordão, SP, Brasil, 2004. Disponível em <http://lbdwww.epfl.ch/~stefano/geoinfolast.pdf>. Acesso em Julho de 2005.

ScVH04 Schilder F., Versley, Y., Habel, C. Extracting Spatial Information: Grounding, Classifying and Linking Spatial Expressions. In *Proceedings of the* A*CM SIGIR Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004. Disponível em <http://www.geo.unizh.ch/~rsp/gir/abstracts/schilder.pdf>. Acesso em Maio de 2005.

SDBD04 Souza, L. A., Delboni, T. M., Borges, K. A. V., Davis Jr., C. A., Laender, A. H. F. Locus: um Localizador Espacial Urbano. In *Proceedings of the VI Brazilian Symposium on Geoinformatics*, Campos do Jordão, SP, Brasil, 2004. Disponível em <http://www.geoinfo.info/geoinfo2004/papers/6400.pdf> Acesso em Fevereiro de 2005.

SDBD05 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. In *Proceedings of the 3rd Latin American Web Congress*, 157-165, Buenos Aires, Argentina, 2005.

SEAS02 Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., Wenke, D. *Ontoedit: Collaborative Ontology Development for the Semantic Web*. In *Proceedings of the 1st International Semantic Web Conference (ISWC2002)*, Vol. 2342 of LNCS, 348-363, Springer-Verlag, 2002. Disponível em <http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2002_iswc_ontoedit.pdf>. Acesso em março de 2004.

Simo87 Simons, P. Parts. A Study in Ontology. Oxford: Clarendon Press, 1987

SMCC06 Silva, M. J., Martins, B., Chaves, M. S., Cardoso, N., Afonso, A. P. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems*, 30 (4): 378-379, 2006.

Smit96 Smith, B. Mereotopology: A Theory of Parts and Boundaries. *Data Knowledge Engineering*, 20 (3): 287-303, 1996.

SmMa98 Smith, B., Mark, D. M. Ontology and Geographic Kinds. In *Proceedings of the International Symposium on Spatial Data Handling*, 308-320, Vancouver, Canada, 1998. Disponível em <http://www.ncgia.buffalo.edu/i21/SDH98.html>. Acesso em junho de 2002.

SmSm77 Smith, J. M., Smith, D. C. P. Database abstractions: aggregation and generalization. *ACM Transactions on Database Systems*, 2 (2): 105-133, 1977.

Sode99 Soderlan, S. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34 (1-3): 233-272, 1999.

Souz05 Souza, L. A., LOCUS: Um Sistema de Localização Geográfica Através de Referências Espaciais Indiretas, Dissertação de mestrado, Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 2005.

SPKR96 **S**wartout, B., Patil, R., Knight, K., Russ, T. *Ontosaurus: A Tool for Browsing and Editing Ontologies*, Gaines, B.R. and Musen, M.A. (Eds.). In *Proceedings of Tenth Knowledge Acquisition Workshop, 69-1 - 69-12*, 1996. Disponível em <http://www.isi.edu/isd/banff_paper/ontosaurus_demo.html>. Acesso em março de 2005.

SwTa99 Swartout, W., Tate, A. Guest Editors' Introduction: Ontologies. *IEEE Intelligent Systems*, 14 (1): 18-19, 1999.

TsLo82 Tsichritzis, D. C., Lochovsky, F. H. *Data Models*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

USCB00 U.S. Census Bureau. Disponível em <http://www.census.gov/geo/www/gazetteer/gazette.html>. Acesso em dezembro de 2005.

Varz03 Varzi, A. Mereology, in Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*, Stanford: CSLI (internet publication), 2003. Disponível em <http://ontology.buffalo.edu/smith/courses03/tb/mereology.pdf>. Acesso em outubro de 2005.

Varz96 Varzi, A. C. Parts, Wholes, and Part-Role Relations: The Prospects of Mereotopology. *Data and Knowledge Engineering*, 20 (3): 259-286, 1996.

VJJS05 Vaid, S., Jones, C. B., Joho, H., Sanderson, M. Spatio-textual Indexing for Geographical Search on the Web. In *Proceedings of 9th International Symposim on Spatial and Temporal Databases*, 218-235, Angra dos Reis, Brasil, 2005.

W3CS01 W3C Semantic Web. Disponível em <http://www.w3.org/2001/sw>. Acesso em Julho de 2003.

WaCS96 Wall, L., Christiansen T., Schwartz, R. L. *Programming Perl*. 2nd Edition, CA: O'Reilly & Associates Inc., 1996.

WKLW98 Weibel, S., Kunze, J., Lagoze, C., Wolf, M. Dublin Core Metadata for Resource Discovery, RFC 2413, The Internet Society, 1998. Disponível em <http://www.ietf.org/rfc/rfc2413.txt>. Acesso em março de 2005.

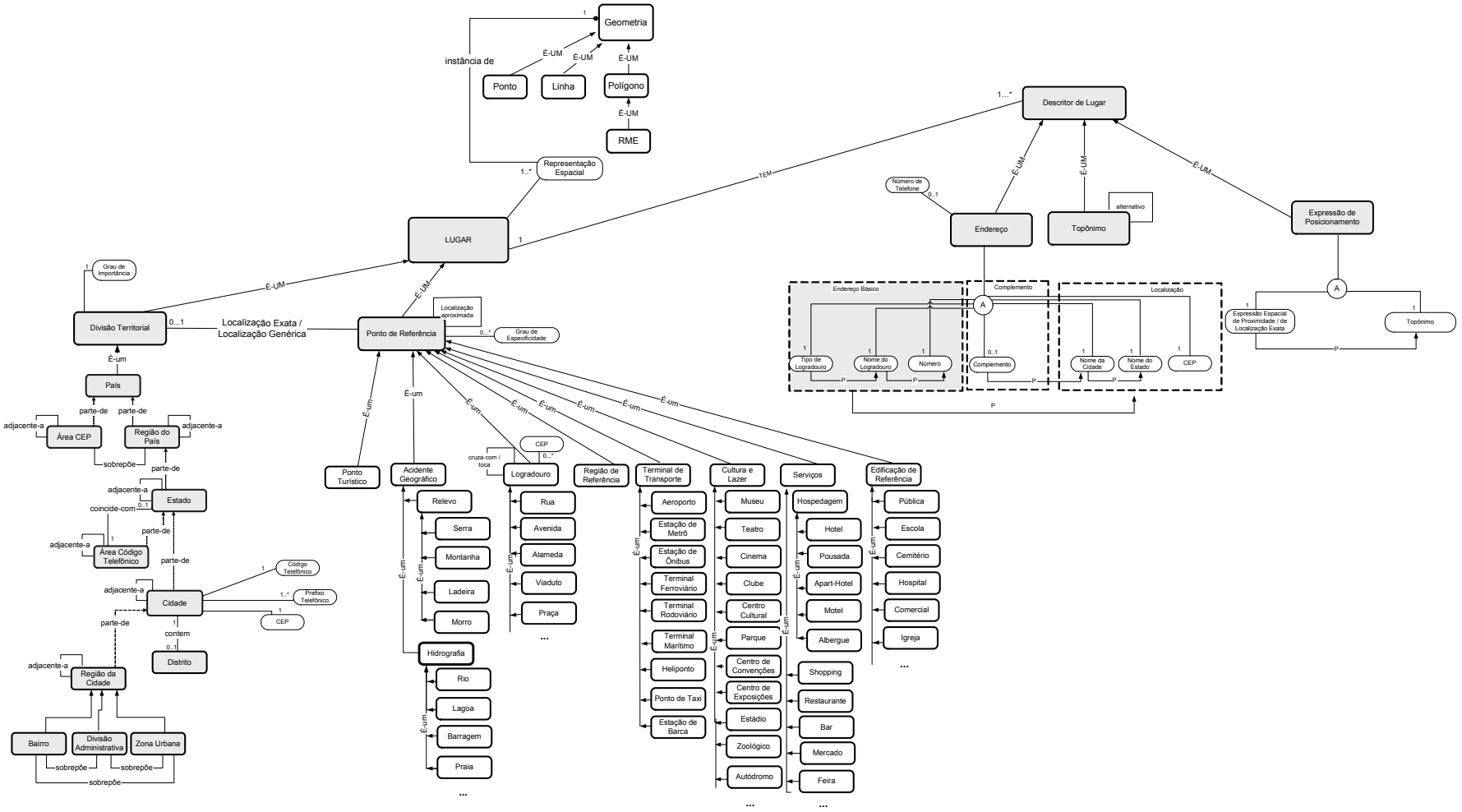
WXWL05 Wang, C., Xie, X., Wang, L., Lu, Y., Ma, W. Detecting Geographic Locations from Web Resources. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval held in conjunction with ACM CIKM 2005*, 17-24, Bremen, Germany, 2005.

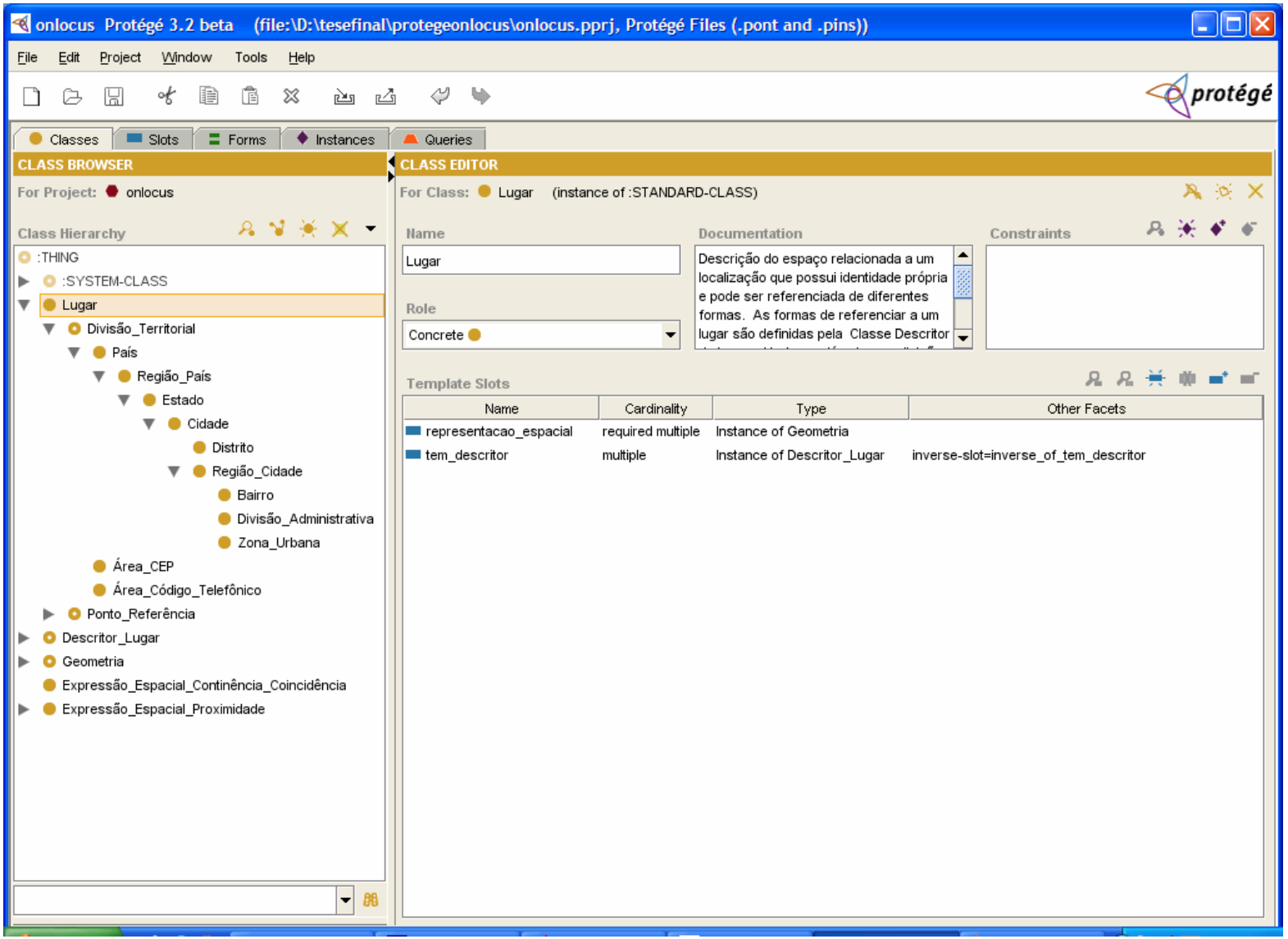
Zhan03 Zhang, Z. *Ontology Query Language for the Semantic Web: A Performance Evaluation.* Mater thesis, Faculty of the Graduate School, University of Georgia, 2005. Disponível em <http://chief.cs.uga.edu/~jam/home/theses/zhijun_thesis/final/zhang_zhijun_200508_ms.pdf>. Acesso em julho de 2005.

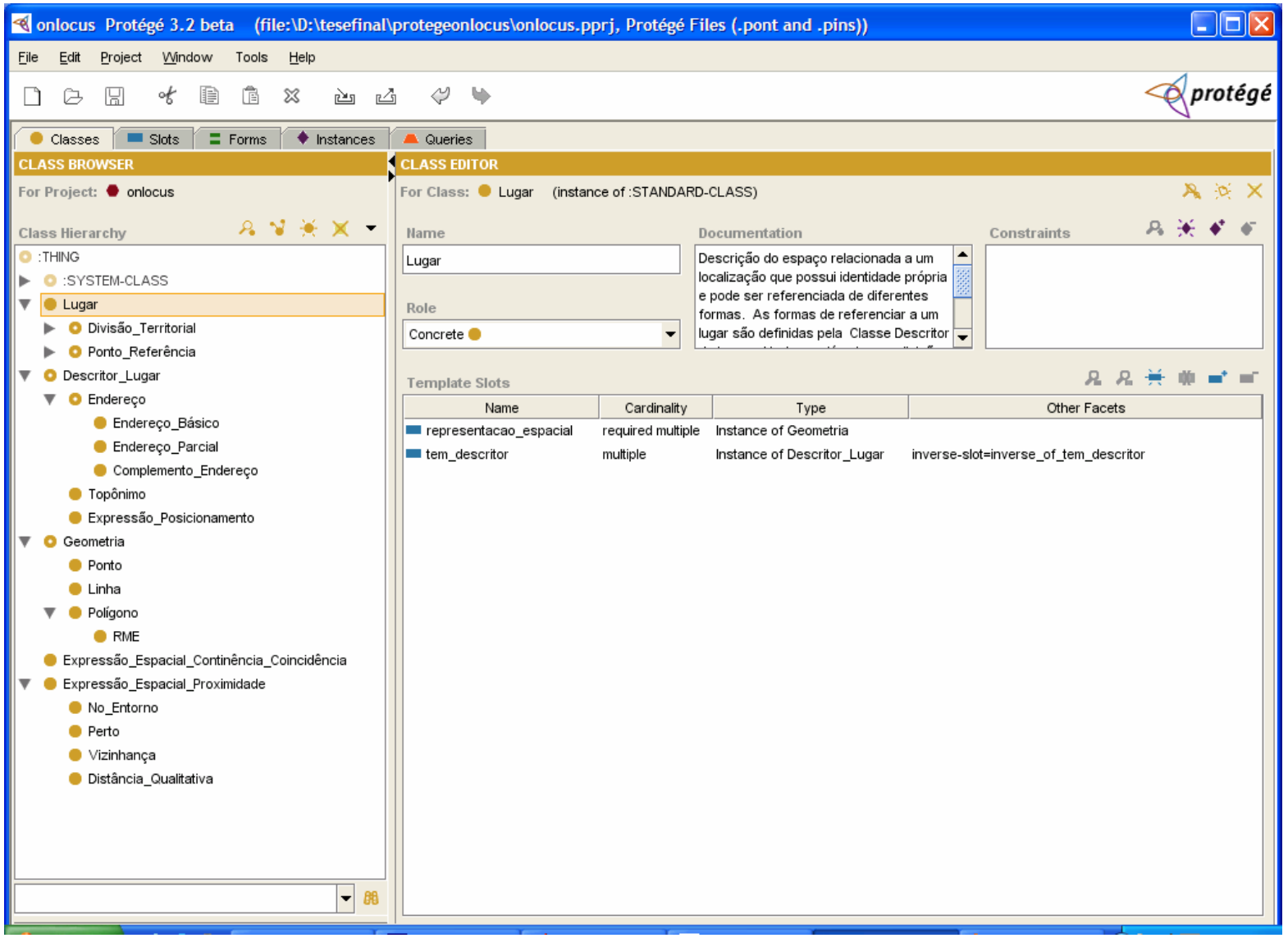
Zhan03a Zhan, F., B. Cognitive Evidence of Vagueness Associated with some Linguistic Terms in Descriptions of Spatial Relations. *Spatial Cognition and Computation*, 3 (2,3): 241-257, 2003.

ZWSL05 Zong, W., Wu, D., Sun, A., Lim, E., Goh, D. H. G. On Assigning Place Names to Geographic Related Web Pages. In *Proceedings of the 5th ACM/IEEE-CS joint Conference on Digital libraries*, 354-362, Denver, CO, USA, 2005.

Anexo A – Ontologia *OnLocus*




onlocus Protégé 3.2 beta (file:\D:\tesefinal\protegeonlocus\onlocus.pprj, Protégé Files (.pont and .pins))
File Edit Project Window Tools Help
protégé
Classes Slots Forms Instances Queries
CLASS BROWSER
For Project: onlocus
Class Hierarchy
:THING
:SYSTEM-CLASS
Lugar
Divisão_Territorial
País
Região_País
Estado
Cidade
Distrito
Região_Cidade
Bairro
Divisão_Administrativa
Zona_Urbana
Área_CEP
Área_Código_Telefônico
Ponto_Referência
Descritor_Lugar
Geometria
Expressão_Espacial_Continência_Coincidência
Expressão_Espacial_Proximidade
CLASS EDITOR
For Class: Lugar (instance of :STANDARD-CLASS)
Name
Lugar
Role
Concrete
Documentation
Descrição do espaço relacionada a um localização que possui identidade própria e pode ser referenciada de diferentes formas. As formas de referenciar a um lugar são definidas pela Classe Descritor
Constraints
Template Slots
Name
Cardinality
Type
Other Facets
representacao_espacial
required multiple
Instance of Geometria
tem_descritor
multiple
Instance of Descritor_Lugar
inverse-slot=inverse_of_tem_descritor


onlocus Protégé 3.2 beta (file:\D:\tesefinal\protegeonlocus\onlocus.pprj, Protégé Files (.pont and .pins))
File Edit Project Window Tools Help
protégé
Classes Slots Forms Instances Queries
CLASS BROWSER
For Project: onlocus
Class Hierarchy
:THING
:SYSTEM-CLASS
Lugar
Divisão_Territorial
Ponto_Referência
Descritor_Lugar
Endereço
Endereço_Básico
Endereço_Parcial
Complemento_Endereço
Topônimo
Expressão_Posicionamento
Geometria
Ponto
Linha
Polígono
RME
Expressão_Espacial_Continência_Coincidência
Expressão_Espacial_Proximidade
No_Entorno
Perto
Vizinhança
Distância_Qualitativa
CLASS EDITOR
For Class: Lugar (instance of :STANDARD-CLASS)
Name
Lugar
Documentation
Descrição do espaço relacionada a um localização que possui identidade própria e pode ser referenciada de diferentes formas. As formas de referenciar a um lugar são definidas pela Classe Descritor
Constraints
Role
Concrete
Template Slots
Name Cardinality Type Other Facets
representacao_espacial required multiple Instance of Geometria
tem_descritor multiple Instance of Descritor_Lugar inverse-slot=inverse_of_tem_descritor

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE rdf:RDF [
        <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
        <!ENTITY rdf_ 'http://protege.stanford.edu/rdf'>
        <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
]>
<rdf:RDF xmlns:rdf="&rdf;"
        xmlns:rdf_="&rdf_;"
        xmlns:rdfs="&rdfs;">
<rdfs:Class rdf:about="&rdf_;Acidente_Geográfico"
        rdfs:label="Acidente_Geográfico">
       <rdfs:comment>Detalhe ou característica natural da superfície do terreno
utilizado como referência de orientação pela população local</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Aeroporto"
        rdfs:label="Aeroporto">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Alameda"
        rdfs:label="Alameda">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Albergue"
        rdfs:label="Albergue">
       <rdfs:subClassOf rdf:resource="&rdf_;Hospedagem"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Apart-hotel"
        rdfs:label="Apart-hotel">
       <rdfs:subClassOf rdf:resource="&rdf_;Hospedagem"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Autódromo"
        rdfs:label="Autódromo">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Avenida"
        rdfs:label="Avenida">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Bairro"
        rdfs:comment="Subdivisão espacial da cidade"
        rdfs:label="Bairro">
       <rdfs:subClassOf rdf:resource="&rdf_;Região_Cidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Bar"
        rdfs:label="Bar">
       <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Bar_restaurante"
        rdfs:label="Bar_restaurante">
       <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Barragem"
        rdfs:label="Barragem">
       <rdfs:subClassOf rdf:resource="&rdf_;Hidrografia"/>
</rdfs:Class>
<rdf:Property rdf:about="&rdf_;CEP"
        rdfs:label="CEP">
       <rdfs:domain rdf:resource="&rdf_;Cidade"/>
       <rdfs:domain rdf:resource="&rdf_;Endereço_Parcial"/>
       <rdfs:domain rdf:resource="&rdf_;Logradouro"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdfs:Class rdf:about="&rdf_;Cemitério"
        rdfs:label="Cemitério">
       <rdfs:subClassOf rdf:resource="&rdf_;Edificação_de_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Centro_Convenções"
```

```
        rdfs:label="Centro_Convenções">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Centro_Cultural"
        rdfs:label="Centro_Cultural">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Centro_Exposições"
        rdfs:label="Centro_Exposições">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Cidade"
        rdfs:comment="Subdivisão política de Estado"
        rdfs:label="Cidade">
       <rdfs:subClassOf rdf:resource="&rdf_;Estado"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Cinema"
        rdfs:label="Cinema">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Clube"
        rdfs:label="Clube">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Complemento_Endereço"
        rdfs:label="Complemento_Endereço">
       <rdfs:comment>parte  opcional  de  um  endereço  que  inclui  dados
complementares do imóvel e o nome do bairro</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Endereço"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Cultura_e_Lazer"
        rdfs:label="Cultura_e_Lazer">
       <rdfs:comment>Lcais  relacionados  a  atividades  culturais  e  de  diversão
usados como referência de orientação  pela população</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Descritor_Lugar"
        rdfs:label="Descritor_Lugar">
       <rdfs:comment>Formas  de  se  referenciar  um  lugar.  Pode  ser  um  endereço,
um topônimo ou uma expressão de posicionamento</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Distrito"
        rdfs:comment="Divisão administrativa da cidade"
        rdfs:label="Distrito">
       <rdfs:subClassOf rdf:resource="&rdf_;Cidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Distância_Qualitativa"
        rdfs:label="Distância_Qualitativa">
       <rdfs:comment>relacionamento  métrico  qualitativo  como,  a  "N"  minutos
de, a "N" Km de</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Expressão_Espacial_Proximidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Divisão_Administrativa"
        rdfs:comment="Subdivisão de uma cidade em regiões administrativas"
        rdfs:label="Divisão_Administrativa">
       <rdfs:subClassOf rdf:resource="&rdf_;Região_Cidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Divisão_Territorial"
        rdfs:label="Divisão_Territorial">
       <rdfs:comment>Representa  o  espaço  territorial  de  um  país  com  suas
diferentes subdivisões geopolíticas, como estado e político-administrativas,
como divisões regionais</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Lugar"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Edificação_Comercial"
        rdfs:label="Edificação_Comercial">
       <rdfs:subClassOf rdf:resource="&rdf_;Edificação_de_Referência"/>
```

```
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Edificação_Pública"
        rdfs:label="Edificação_Pública">
       <rdfs:subClassOf rdf:resource="&rdf_;Edificação_de_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Edificação_de_Referência"
        rdfs:label="Edificação_de_Referência">
       <rdfs:comment>Construções  utilizadas como referência de orientação pela
população local</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Elevado"
        rdfs:label="Elevado">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro_Obra_de_Arte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Endereço"
        rdfs:label="Endereço">
       <rdfs:comment>Forma de expressar a localização de um lugar de interesse.
O Endereço Básico precede o Complemento e o Complemento precede o Endereço
Parcial</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Descritor_Lugar"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Endereço_Básico"
        rdfs:comment="parte do endereço  que identifica o imóvel"
        rdfs:label="Endereço_Básico">
       <rdfs:subClassOf rdf:resource="&rdf_;Endereço"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Endereço_Parcial"
        rdfs:comment="Parte do endereço que identifica sua localização"
        rdfs:label="Endereço_Parcial">
       <rdfs:subClassOf rdf:resource="&rdf_;Endereço"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Escola"
        rdfs:label="Escola">
       <rdfs:subClassOf rdf:resource="&rdf_;Edificação_de_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Estado"
        rdfs:comment="Subdivisão política do País"
        rdfs:label="Estado">
       <rdfs:subClassOf rdf:resource="&rdf_;Região_País"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Estação_Barca"
        rdfs:label="Estação_Barca">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Estação_Metrô"
        rdfs:label="Estação_Metrô">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Estação_Ônibus"
        rdfs:label="Estação_Ônibus">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Estádio_Futebol"
        rdfs:label="Estádio_Futebol">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdf:Property rdf:about="&rdf_;Expressão"
        rdfs:label="Expressão">
       <rdfs:comment>expressão  que  denota  um  relacionamento  espacial  de
continência, coincidência e proximidade, em linguagem natural</rdfs:comment>
       <rdfs:domain rdf:resource="&rdf_;Expressão_Posicionamento"/>
       <rdfs:range rdf:resource="&rdfs;Resource"/>
</rdf:Property>
<rdfs:Class rdf:about="&rdf_;Expressão_Espacial_Continência_Coincidência"
        rdfs:label="Expressão_Espacial_Continência_Coincidência">
```

```
      <rdfs:comment>Expressões espaciais de continência e coincidência em
linguagem natural. Corresponde aos relacionamentos topológicos: "está contido
em" e "coincide com"</rdfs:comment>
      <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Expressão_Espacial_Proximidade"
       rdfs:label="Expressão_Espacial_Proximidade">
      <rdfs:comment>Expressões em linguagem natural que denotam proximidade e
são utilizadas em Expressões de Posicionamento</rdfs:comment>
      <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Expressão_Posicionamento"
       rdfs:label="Expressão_Posicionamento">
      <rdfs:comment  xml:space='preserve'><![CDATA[Construção  semântica  em
linguagem natural utilizada para expressar, por meio de um par <expressão que
denota um relacionamento espacial,ponto de referência>, a posição relativa no
espaço de um lugar de interesse.]]></rdfs:comment>
      <rdfs:subClassOf rdf:resource="&rdf_;Descritor_Lugar"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Feira"
       rdfs:label="Feira">
      <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Geometria"
       rdfs:comment="Descrição geométrica de um lugar"
       rdfs:label="Geometria">
      <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ginásio_Poliesportivo"
       rdfs:label="Ginásio_Poliesportivo">
      <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Heliponto"
       rdfs:label="Heliponto">
      <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Hidrografia"
       rdfs:comment="elementos hidrográficos permanentes ou temporários"
       rdfs:label="Hidrografia">
      <rdfs:subClassOf rdf:resource="&rdf_;Acidente_Geográfico"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Hospedagem"
       rdfs:label="Hospedagem">
      <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Hospital"
       rdfs:label="Hospital">
      <rdfs:subClassOf rdf:resource="&rdf_;Edificação_de_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Hotel"
       rdfs:label="Hotel">
      <rdfs:subClassOf rdf:resource="&rdf_;Hospedagem"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Igreja"
       rdfs:label="Igreja">
      <rdfs:subClassOf rdf:resource="&rdf_;Edificação_de_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ladeira"
       rdfs:label="Ladeira">
      <rdfs:subClassOf rdf:resource="&rdf_;Relevo"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Lagoa"
       rdfs:label="Lagoa">
      <rdfs:subClassOf rdf:resource="&rdf_;Hidrografia"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Linha"
       rdfs:label="Linha">
```

```
      <rdfs:comment>poligonais - segmentos de reta compreendidos entre dois
pontos</rdfs:comment>
      <rdfs:subClassOf rdf:resource="&rdf_;Geometria"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Logradouro"
       rdfs:label="Logradouro">
      <rdfs:comment>Espaço público de uso comum destinado ao trânsito de
pedestre e/ou veículos usado como referência pela população</rdfs:comment>
      <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Logradouro_Obra_de_Arte"
       rdfs:label="Logradouro_Obra_de_Arte">
      <rdfs:comment>logradouro que exige maiores conhecimentos de engenharia
para ser criado como, viaduto, túnel, trincheira</rdfs:comment>
      <rdfs:subClassOf rdf:resource="&rdf_;Logradouro"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Lugar"
       rdfs:label="Lugar">
      <rdfs:comment>Descrição do espaço relacionada a um localização que
possui identidade própria e pode ser referenciada de diferentes formas.  As
formas de referenciar a um lugar são definidas pela  Classe Descritor de
Lugar.  Um lugar além de uma divisão territorial ou ponto de referência pode
ser qualquer espaço ou objeto identificado por um dos descritores de lugar
definidos</rdfs:comment>

      <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Mercado"
       rdfs:label="Mercado">
      <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Montanha"
       rdfs:label="Montanha">
      <rdfs:subClassOf rdf:resource="&rdf_;Relevo"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Morro"
       rdfs:label="Morro">
      <rdfs:subClassOf rdf:resource="&rdf_;Relevo"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Motel"
       rdfs:label="Motel">
      <rdfs:subClassOf rdf:resource="&rdf_;Hospedagem"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Museu"
       rdfs:label="Museu">
      <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;No_Entorno"
       rdfs:comment="muito próximo- em frente, do lado, atrás"
       rdfs:label="No_Entorno">
      <rdfs:subClassOf rdf:resource="&rdf_;Expressão_Espacial_Proximidade"/>
</rdfs:Class>
<rdf:Property rdf:about="&rdf_;Nome_cidade"
       rdfs:label="Nome_cidade">
      <rdfs:range rdf:resource="&rdf_;Cidade"/>
      <rdfs:domain rdf:resource="&rdf_;Endereço_Parcial"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;Nome_estado"
       rdfs:comment="nome do estado de onde é o endereço"
       rdfs:label="Nome_estado">
      <rdfs:domain rdf:resource="&rdf_;Endereço_Parcial"/>
      <rdfs:range rdf:resource="&rdf_;Estado"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;Número_telefone"
       rdfs:label="Número_telefone">
      <rdfs:domain rdf:resource="&rdf_;Endereço_Parcial"/>
      <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
```

```
<rdfs:Class rdf:about="&rdf_;Parque"
        rdfs:label="Parque">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;País"
        rdfs:comment="Nação"
        rdfs:label="País">
       <rdfs:subClassOf rdf:resource="&rdf_;Divisão_Territorial"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Perto"
        rdfs:label="Perto">
       <rdfs:comment>um pouco mais longe do que no entorno como, perto de, a
poucos minutos de</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Expressão_Espacial_Proximidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Polígono"
        rdfs:label="Polígono">
       <rdfs:comment>região do plano limitada por uma linha poligonal fechada.
o primeiro ponto coincide com o último.</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Geometria"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ponte"
        rdfs:label="Ponte">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro_Obra_de_Arte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ponto"
        rdfs:comment="par ordenado (x,y) de coordenadas espaciais"
        rdfs:label="Ponto">
       <rdfs:subClassOf rdf:resource="&rdf_;Geometria"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ponto_Referência"
        rdfs:label="Ponto_Referência">
       <rdfs:comment>Representa um local, um marco urbano ou um marco ambiental
reconhecido pelas pessoas e que possa ser usado para orientação
espacial.</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Lugar"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ponto_Turístico"
        rdfs:comment="Locais de interesse turístico reconhecido pela
população."
        rdfs:label="Ponto_Turístico">
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Ponto_taxi"
        rdfs:label="Ponto_taxi">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Pousada"
        rdfs:label="Pousada">
       <rdfs:subClassOf rdf:resource="&rdf_;Hospedagem"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Praia"
        rdfs:label="Praia">
       <rdfs:subClassOf rdf:resource="&rdf_;Hidrografia"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Praça"
        rdfs:label="Praça">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;RME"
        rdfs:comment="Retângulo Mínimo Envolvente"
        rdfs:label="RME">
       <rdfs:subClassOf rdf:resource="&rdf_;Polígono"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Região_Cidade"
        rdfs:comment="Subdivisão territorial de uma cidade"
        rdfs:label="Região_Cidade">
       <rdfs:subClassOf rdf:resource="&rdf_;Cidade"/>
```

```
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Região_País"
        rdfs:label="Região_País">
       <rdfs:comment>Representa as subdivisões  territoriais de um País segundo
diferentes critérios. Exemplo: mesorregiões, regiões geográficas (norte,
sul..)</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;País"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Região_de_Referência"
        rdfs:label="Região_de_Referência">
       <rdfs:comment>Nome de uma região usada como referência de localização
pela população local</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Relevo"
        rdfs:comment="qualquer aspecto relativo ao modelado do terreno"
        rdfs:label="Relevo">
       <rdfs:subClassOf rdf:resource="&rdf_;Acidente_Geográfico"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Restaurante"
        rdfs:label="Restaurante">
       <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Rio"
        rdfs:label="Rio">
       <rdfs:subClassOf rdf:resource="&rdf_;Hidrografia"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Rua"
        rdfs:label="Rua">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Serra"
        rdfs:label="Serra">
       <rdfs:subClassOf rdf:resource="&rdf_;Relevo"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Serviços"
        rdfs:label="Serviços">
       <rdfs:comment>Atividades comerciais relacionadas à prestação de serviço
usadas como referência pela população</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Shopping"
        rdfs:label="Shopping">
       <rdfs:subClassOf rdf:resource="&rdf_;Serviços"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Teatro"
        rdfs:label="Teatro">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Terminal_Ferroviário"
        rdfs:label="Terminal_Ferroviário">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Terminal_Marítimo"
        rdfs:label="Terminal_Marítimo">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Terminal_Rodoviário"
        rdfs:label="Terminal_Rodoviário">
       <rdfs:subClassOf rdf:resource="&rdf_;Terminal__de_Transporte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Terminal__de_Transporte"
        rdfs:label="Terminal _de_Transporte">
       <rdfs:comment>Estações de embarque e desembarque de passageiros usadas
como referência de localização pela população</rdfs:comment>
       <rdfs:subClassOf rdf:resource="&rdf_;Ponto_Referência"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Topônimo"
```

```
        rdfs:comment="Nomes próprios de lugares possuindo nomes alternativos"
        rdfs:label="Topônimo">
       <rdfs:subClassOf rdf:resource="&rdf_;Descritor_Lugar"/>
</rdfs:Class>
<rdf:Property rdf:about="&rdf_;Topônimo_expressão"
        rdfs:comment="nome do ponto de referência"
        rdfs:label="Topônimo_expressão">
       <rdfs:domain rdf:resource="&rdf_;Expressão_Posicionamento"/>
       <rdfs:range rdf:resource="&rdfs;Class"/>
</rdf:Property>
<rdfs:Class rdf:about="&rdf_;Trincheira"
        rdfs:label="Trincheira">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro_Obra_de_Arte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Túnel"
        rdfs:label="Túnel">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro_Obra_de_Arte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Viaduto"
        rdfs:label="Viaduto">
       <rdfs:subClassOf rdf:resource="&rdf_;Logradouro_Obra_de_Arte"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Vizinhança"
        rdfs:comment="nos arredores"
        rdfs:label="Vizinhança">
       <rdfs:subClassOf rdf:resource="&rdf_;Expressão_Espacial_Proximidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Zona_Urbana"
        rdfs:comment="Subdivisão regional da cidade - zona norte,    zona
sul...."
        rdfs:label="Zona_Urbana">
       <rdfs:subClassOf rdf:resource="&rdf_;Região_Cidade"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Zoológico"
        rdfs:label="Zoológico">
       <rdfs:subClassOf rdf:resource="&rdf_;Cultura_e_Lazer"/>
</rdfs:Class>
<rdf:Property rdf:about="&rdf_;adjacente_a_Cod_telefônico"
        rdfs:label="adjacente_a_Cod_telefônico">
       <rdfs:domain rdf:resource="&rdf_;Área_Código_Telefônico"/>
       <rdfs:range rdf:resource="&rdfs;Class"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;adjacente_a_cidade"
        rdfs:label="adjacente_a_cidade">
       <rdfs:range rdf:resource="&rdf_;Cidade"/>
       <rdfs:domain rdf:resource="&rdf_;Cidade"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;adjacente_a_estado"
        rdfs:label="adjacente_a_estado">
       <rdfs:domain rdf:resource="&rdf_;Estado"/>
       <rdfs:range rdf:resource="&rdf_;Estado"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;adjacente_a_região"
        rdfs:label="adjacente_a_região">
       <rdfs:domain rdf:resource="&rdf_;Região_País"/>
       <rdfs:range rdf:resource="&rdf_;Região_País"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;adjacente_região_cidade"
        rdfs:label="adjacente_região_cidade">
       <rdfs:domain rdf:resource="&rdf_;Cidade"/>
       <rdfs:range rdf:resource="&rdf_;Região_Cidade"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;adjacente_área_CEP"
        rdfs:label="adjacente_área_CEP">
       <rdfs:domain rdf:resource="&rdf_;Área_CEP"/>
       <rdfs:range rdf:resource="&rdfs;Class"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;cod_área"
```

```
        rdfs:comment="DDD"
        rdfs:label="cod_área">
       <rdfs:domain rdf:resource="&rdf_;Área_Código_Telefônico"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;coincide_com"
        rdfs:label="coincide_com">
       <rdfs:range rdf:resource="&rdf_;Estado"/>
       <rdfs:domain rdf:resource="&rdf_;Área_Código_Telefônico"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;complemento"
        rdfs:label="complemento">
       <rdfs:comment>inclui dados complementares de um endereço como número do
apartamento, sala, nome do bairro</rdfs:comment>
       <rdfs:domain rdf:resource="&rdf_;Complemento_Endereço"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;coordenada1_y"
        rdfs:label="coordenada1_y">
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;coordenada2_Y"
        rdfs:label="coordenada2_Y">
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;coordenada2_x"
        rdfs:label="coordenada2_x">
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;coordenada_x"
        rdfs:label="coordenada_x">
       <rdfs:domain rdf:resource="&rdf_;Ponto"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;coordenada_y"
        rdfs:label="coordenada_y">
       <rdfs:domain rdf:resource="&rdf_;Ponto"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;código_telefônico"
        rdfs:label="código_telefônico">
       <rdfs:domain rdf:resource="&rdf_;Cidade"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;distância_quantificada"
        rdfs:label="distância_quantificada">
       <rdfs:comment>área em volta do ponto de referência cujo raio corresponde
ao raio do RME somado ao valor informado</rdfs:comment>
       <rdfs:domain rdf:resource="&rdf_;Expressão_Espacial_Proximidade"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;expressão_continência_coincidência"
        rdfs:comment="expressão de continência e coincidência em linguagem
natural"
        rdfs:label="expressão_continência_coincidência">
       <rdfs:domain
rdf:resource="&rdf_;Expressão_Espacial_Continência_Coincidência"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;expresão_proximidade"
        rdfs:label="expresão_proximidade">
       <rdfs:domain rdf:resource="&rdf_;Expressão_Espacial_Proximidade"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;grau_especificidade"
        rdfs:label="grau_especificidade">
       <rdfs:comment>informa a probabilidade de uma instância de um ponto de
referência ser única na cidade, no estado e no país</rdfs:comment>
```

```
      <rdfs:domain rdf:resource="&rdf_;Ponto_Referência"/>
      <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_coincide_com"
       rdfs:label="inverse_of_coincide_com">
      <rdfs:domain rdf:resource="&rdf_;Estado"/>
      <rdfs:range rdf:resource="&rdf_;Área_Código_Telefônico"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_localização_exata"
       rdfs:label="inverse_of_localização_exata">
      <rdfs:domain rdf:resource="&rdf_;Divisão_Territorial"/>
      <rdfs:domain rdf:resource="&rdf_;Endereço"/>
      <rdfs:range rdf:resource="&rdf_;Ponto_Referência"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_localização_genérica"
       rdfs:label="inverse_of_localização_genérica">
      <rdfs:domain rdf:resource="&rdf_;Divisão_Territorial"/>
      <rdfs:range rdf:resource="&rdf_;Ponto_Referência"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_pertence_a"
       rdfs:label="inverse_of_pertence_a">
      <rdfs:domain rdf:resource="&rdf_;Divisão_Territorial"/>
      <rdfs:range rdf:resource="&rdf_;Endereço"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_sobrepõe"
       rdfs:label="inverse_of_sobrepõe">
      <rdfs:range rdf:resource="&rdf_;Bairro"/>
      <rdfs:domain rdf:resource="&rdf_;Divisão_Administrativa"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_sobrepõe_bairro"
       rdfs:label="inverse_of_sobrepõe_bairro">
      <rdfs:domain rdf:resource="&rdf_;Bairro"/>
      <rdfs:range rdf:resource="&rdf_;Zona_Urbana"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_sobrepõe_região_país"
       rdfs:label="inverse_of_sobrepõe_região_país">
      <rdfs:domain rdf:resource="&rdf_;Região_País"/>
      <rdfs:range rdf:resource="&rdf_;Área_CEP"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_sobrepõe_zona_urbana"
       rdfs:label="inverse_of_sobrepõe_zona_urbana">
      <rdfs:range rdf:resource="&rdf_;Divisão_Administrativa"/>
      <rdfs:domain rdf:resource="&rdf_;Zona_Urbana"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;inverse_of_tem_descritor"
       rdfs:label="inverse_of_tem_descritor">
      <rdfs:domain rdf:resource="&rdf_;Descritor_Lugar"/>
      <rdfs:range rdf:resource="&rdf_;Lugar"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;localização_aproximada"
       rdfs:label="localização_aproximada">
      <rdfs:comment>um objeto espacial é localizado em função da localização
de       outro.       Forma       de       Expressão       definida       em
Expressão_Espacial_Proximidade</rdfs:comment>
      <rdfs:domain rdf:resource="&rdf_;Ponto_Referência"/>
      <rdfs:range rdf:resource="&rdf_;Ponto_Referência"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;localização_exata"
       rdfs:label="localização_exata">
      <rdfs:comment>Endereço onde está localizado um ponto de referência.
Forma             de             expressão             definida             em
Expressão_Espacial_Continência_Coincidência</rdfs:comment>
      <rdfs:domain rdf:resource="&rdf_;Ponto_Referência"/>
      <rdfs:range rdf:resource="&rdfs;Resource"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;localização_genérica"
       rdfs:label="localização_genérica">
```

```
      <rdfs:comment>relacionamento binário entre objetos espaciais e regiões
hierárquicas do espaço usado quando a localização exata ou aproximada não foi
possível. Não existe um comprometimento com a localização precisa significando
que um objeto X está genericamente localizado em alguma parte da região Y.
Forma de expressão "localizado_em"</rdfs:comment>
       <rdfs:range rdf:resource="&rdf_;Divisão_Territorial"/>
       <rdfs:domain rdf:resource="&rdf_;Ponto_Referência"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;nome"
        rdfs:comment="nome de lugar"
        rdfs:label="nome">
       <rdfs:domain rdf:resource="&rdf_;Topônimo"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;nome_logradouro"
        rdfs:label="nome_logradouro">
       <rdfs:domain rdf:resource="&rdf_;Endereço_Básico"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;nome_sede"
        rdfs:comment="nome da cidade sede da região postal"
        rdfs:label="nome_sede">
       <rdfs:domain rdf:resource="&rdf_;Área_CEP"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;número_imóvel"
        rdfs:label="número_imóvel">
       <rdfs:domain rdf:resource="&rdf_;Endereço_Básico"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;onlocus_Slot_11"
        rdfs:label="onlocus_Slot_11">
       <rdfs:domain rdf:resource="&rdf_;Bairro"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;onlocus_Slot_3"
        rdfs:label="onlocus_Slot_3">
       <rdfs:domain rdf:resource="&rdf_;Estado"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;onlocus_Slot_8"
        rdfs:label="onlocus_Slot_8">
       <rdfs:domain rdf:resource="&rdf_;Cidade"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;par_coordenadas_XY"
        rdfs:label="par_coordenadas_XY">
       <rdfs:domain rdf:resource="&rdf_;Linha"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;par_cordenadadasXY"
        rdfs:label="par_cordenadadasXY">
       <rdfs:domain rdf:resource="&rdf_;Polígono"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;pertence_a"
        rdfs:label="pertence_a">
       <rdfs:range rdf:resource="&rdf_;Divisão_Territorial"/>
       <rdfs:domain rdf:resource="&rdf_;Endereço"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;possui_nome_alternativo"
        rdfs:label="possui_nome_alternativo">
       <rdfs:domain rdf:resource="&rdf_;Topônimo"/>
       <rdfs:range rdf:resource="&rdf_;Topônimo"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;prefixo_telefônico"
        rdfs:label="prefixo_telefônico">
       <rdfs:domain rdf:resource="&rdf_;Cidade"/>
```

```
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;representacao_espacial"
        rdfs:label="representacao_espacial">
       <rdfs:range rdf:resource="&rdf_;Geometria"/>
       <rdfs:domain rdf:resource="&rdf_;Lugar"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;sobrepõe"
        rdfs:label="sobrepõe">
       <rdfs:domain rdf:resource="&rdf_;Bairro"/>
       <rdfs:range rdf:resource="&rdf_;Divisão_Administrativa"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;sobrepõe_bairro"
        rdfs:label="sobrepõe_bairro">
       <rdfs:range rdf:resource="&rdf_;Bairro"/>
       <rdfs:domain rdf:resource="&rdf_;Zona_Urbana"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;sobrepõe_região_país"
        rdfs:label="sobrepõe_região_país">
       <rdfs:range rdf:resource="&rdf_;Região_País"/>
       <rdfs:domain rdf:resource="&rdf_;Área_CEP"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;sobrepõe_zona_urbana"
        rdfs:label="sobrepõe_zona_urbana">
       <rdfs:domain rdf:resource="&rdf_;Divisão_Administrativa"/>
       <rdfs:range rdf:resource="&rdf_;Zona_Urbana"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;tem_descritor"
        rdfs:label="tem_descritor">
       <rdfs:range rdf:resource="&rdf_;Descritor_Lugar"/>
       <rdfs:domain rdf:resource="&rdf_;Lugar"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;tipo_logradouro"
        rdfs:label="tipo_logradouro">
       <rdfs:domain rdf:resource="&rdf_;Endereço_Básico"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdf:Property rdf:about="&rdf_;tipo_região"
        rdfs:label="tipo_região">
       <rdfs:domain rdf:resource="&rdf_;Região_País"/>
       <rdfs:range rdf:resource="&rdfs;Literal"/>
</rdf:Property>
<rdfs:Class rdf:about="&rdf_;Área_CEP"
        rdfs:comment="Regiões postais que subdividem o País"
        rdfs:label="Área_CEP">
       <rdfs:subClassOf rdf:resource="&rdf_;Divisão_Territorial"/>
</rdfs:Class>
<rdfs:Class rdf:about="&rdf_;Área_Código_Telefônico"
        rdfs:comment="Subdivisão do País em áreas de telefonia"
        rdfs:label="Área_Código_Telefônico">
       <rdfs:subClassOf rdf:resource="&rdf_;Divisão_Territorial"/>
</rdfs:Class>
</rdf:RDF>
```

Anexo B – Notação EBNF [48]Utilizada

BNF Extendido	Significado
“.....”	Símbolo terminal
<.....>	Símbolo não terminal
(.....)	Agrupamento de alternativas
[.....]	Símbolos opcionais (0 ou 1 instância)
{.....}	Símbolos que repetem 0 ou mais vezes
{.....}$^{+}$	Símbolos que repetem uma ou mais vezes
{.....}^{n}	Símbolos que repetem uma ou n vezes
::=	Definição do símbolo (é definido como)
;	Fim de produção
|	Alternativa (ou)
,	Concatenação
(*.....*)	Comentário

[48] ISO/IEC 14977:1996(E)

Anexo C – Padrões definidos em PERL

```
#!/usr/bin/perl

############
# Padrões: #
############

# Gerais:
my $espaco = "(?:\\s|<S>)";
my $caracterSeparador = "(?:\\.|,|\\-|\\/|;)";
my $separadorFraco = "(?:$espaco{0,10}(?:$caracterSeparador|$espaco)$espaco{0,10}){1,10}";
my $separadorForte = "(?:$espaco{0,10}$caracterSeparador$espaco{0,10}){1,10}";
my $separadorForteOuTag = "(?:$espaco{0,10}(?:$caracterSeparador|<S>)$espaco{0,10}){1,10}";

# Cidade/Estado:
my $identificadorCidadeEstado = "(?:[Cc]idade:)";
my $preposicao = "(?:de|da|do|das|dos)";
my $nome = "(?:[A-Z][a-z]+$espaco$preposicao$espaco|[A-Z][a-z]+$espaco?)+";
my $siglaEstado = "(?:AC|AL|AM|AP|BA|CE|DF|ES|GO|MA|MG|MS|MT|PA|PB|PE|PI|PR|RJ|RN|RO|RR|RS|SC|SE|SP|TO)";
my $separadorNomeEstadoComposto = "\\.$espaco|\\.|$espaco";
my $nomeEstado = "(?:Acre|ACRE|Alagoas|ALAGOAS|Amazonas|AMAZONAS|Amapa|AMAPA|Bahia|BAHIA|Ceara|CEARA|Distrito
      Federal|DISTRITO
      FEDERAL|D(?:$separadorNomeEstadoComposto)Federal|D(?:$separadorNomeEstadoComposto)FEDERAL|Espirito
      Santo|ESPIRITO
      SANTO|E(?:$separadorNomeEstadoComposto)Santo|E(?:$separadorNomeEstadoComposto)SANTO|Goias|GOIAS|Maranhao|MARAN
      HAO|Minas Gerais|MINAS
      GERAIS|M(?:$separadorNomeEstadoComposto)Gerais|M(?:$separadorNomeEstadoComposto)GERAIS|Mato Grosso do Sul|MATO
      GROSSO DO SUL|Mato Grosso|MATO
      GROSSO|M(?:$separadorNomeEstadoComposto)Grosso|M(?:$separadorNomeEstadoComposto)GROSSO|Paraiba|PARAIBA|Pernamb
      uco|PERNAMBUCO|Piaui|PIAUI|Parana|PARANA|Rio de Janeiro|RIO DE JANEIRO|Rio Grande do Norte|RIO GRANDE DO
```

```
      NORTE|Roraima|RORAIMA|Rondonia|RONDONIA|Rio Grande do Sul|RIO GRANDE DO SUL|Santa Catarina|SANTA
      CATARINA|S(?:$separadorNomeEstadoComposto)Catarina|S(?:$separadorNomeEstadoComposto)CATARINA|Sergipe|SERGIPE|S
      ao Paulo|SÃO
      PAULO|S(?:$separadorNomeEstadoComposto)Paulo|S(?:$separadorNomeEstadoComposto)PAULO|R(?:$separadorNomeEstadoCo
      mposto)de Janeiro|R(?:$separadorNomeEstadoComposto)de JANEIRO|Tocantins|TOCANTINS)";
my $cidadeEstado = "$identificadorCidadeEstado?$separadorFraco(($nome)$separadorFraco($siglaEstado|$nomeEstado)|
      ($nome)$separadorFraco\\(($siglaEstado|$nomeEstado)\\))";
# CEP:
my $identificadorCEP = "(?:CEP|Cep)[\\:|\\.]$espaco*";
my $prefixoCEP = "[0-9][0-9]\\.?[0-9][0-9][0-9]";
my $separadorCEP = "(?:\\-|$espaco+)";
my $sufixoCEP = "[0-9][0-9][0-9]";
my $numeroCEPComSeparador = "$prefixoCEP$espaco*$separadorCEP$espaco*$sufixoCEP";
my $numeroCEPSemSeparador = "[0-9][0-9][0-9][0-9][0-9][0-9][0-9][0-9]";
my $CEP = "($identificadorCEP$separadorFraco($numeroCEPComSeparador|$numeroCEPSemSeparador)|
      ($numeroCEPComSeparador))";

# Telefone:
my $identificadorTelefone =
      "(?:Telefone|TELEFONE|Tel\\.?|TEL\\.?|Fone\\.?|FONE\\.?|Fax\\.?|FAX\\.?|TeleFax|TELEFAX|Fone/Fax|FONE/FAX)$esp
      aco*\\:$espaco*";
my $DDD = "(?:0?[1-9][0-9]|0[Xx][Xx]$espaco*[1-9][0-9]|0\\.[Xx][Xx]\\.[1-9][0-9]|0\\.\\*\\*\\.[1-9][0-9])";
my $prefixoTelefone = "(?:[1-9][0-9][0-9]|[1-9][0-9][0-9][0-9])";
my $separadorTelefone = "(?:$espaco*[\\.\\-]$espaco*|$espaco+)";
my $sufixoTelefone = "[0-9][0-9][0-9][0-9]";
my $telefone = "(($DDD)$espaco+($prefixoTelefone)$separadorTelefone($sufixoTelefone)|" . # 31 3462-6648
      "\\(($DDD)\\)$espaco*($prefixoTelefone)$separadorTelefone($sufixoTelefone))"; # (31) 3462-6648

# Endereco:
my $identificadorEndereco = "(?:Endereco|ENDERECO|End\\.?|END\\.?|Localizacao|Local)$espaco*\\:$espaco*";
my $nomeLogradouro = "(?:(?:$preposicao$espaco+)?$nome|[A-Z]+[0-9]*|(?:[1-9]$espaco+|[1-2][0-9]$espaco+|[3][0-
      1]$espaco+)$preposicao$espaco+[A-Z][a-z]+|(?:[1-9]$espaco+|[0-9]+))";
my $separadorEndereco = "(?:$espaco*,$espaco*|$espaco*\\-$espaco*|$espaco+)";
my $tipoLogradouro =
      "(?:Rua|RUA|R\\.?|Av\\.?|AV\\.?|Ave\\.|AVENIDA|Avenida|ROD\\.?|Rod\\.?|Rodovia|RODOVIA|Pc\\.?|Praca|Alameda|Al
      \\.?|Travessa|Tv\\.?|Largo|LG\\.|Lg\\.|Ladeira|Lad\\.?|Estrada|Est\\.?)";
my $rodovia = "(?:BR [0-9]+|BR\\.$espaco+[0-9]+)";
```

```
my $num = "(?:n\\.?|N\\.?|No\\.|no\\.|Km\\.?|KM\\.?|Numero|numero)";
my $numeroEndereco = "(?:(?:[0-9]{1,3}\\.[0-9]{3}|[0-9]+)|$num$espaco*(?:[0-9]{1,3}\\.[0-9]{3}|[0-
9]+)|$num$espaco*(?:[0-9]{1,3}\\.[0-9]{3}|[0-9]+)$espaco*(?:[a-z]|[A-Z])|s\\/n|S\\/N)";
my $endereco = "($tipoLogradouro)$espaco+($nomeLogradouro)$separadorEndereco($numeroEndereco)";
my $enderecoRodovia = "($rodovia)$separadorEndereco($numeroEndereco)";
my $complementoEndereco = "(.{0,30})";
```