Capítulo 8

Criação e expansão de geo-ontologias, dimensionamento de informação geográfica e reconhecimento de locais e seus relacionamentos em textos

Marcirio Chaves

Luís Costa, Diana Santos e Nuno Cardoso, editores, *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Capítulo 8, p. 49–59, 2008

Este artigo resume o trabalho desenvolvido ao longo de mais de quatro anos na Linguateca no âmbito do meu doutorado. Até 2004, a maior parte das fontes de dados geográficos de Portugal encontrava-se distribuída, desintegrada e desconexa. Essas fontes contêm informação complementar, heterogênea e semi-estruturada. Qualquer aplicação que necessitasse utilizá-las tinha que recorrer a diversos bancos de dados, estudar seus esquemas conceituais e traduzir a informação para um formato comum de representação, entre outras tarefas. Além disso, os dados armazenados em bancos de dados proprietários são invisíveis para aplicações da Web Semântica.

Nesse contexto havia a necessidade da criação de um modelo genérico suficiente para reunir informação geográfica de diversas fontes, de múltiplos domínios geográficos (e.g. administrativo e físico) e disponibilizá-la de forma integrada e em um formato legível por máquina. Assim, foi criada a GKB (*Geographic Knowledge Base*) (Chaves et al., 2005a,b), um sistema de gerenciamento de conhecimento geográfico, ilustrado na figura 8.1 e descrito na próxima seção.

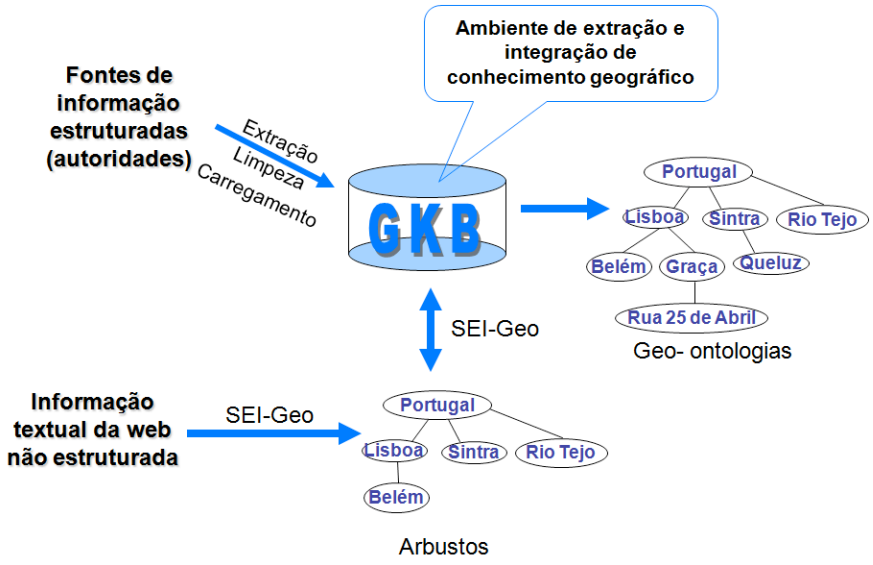

Figura 8.1: Arquitetura global do sistema de gerenciamento de conhecimento geográfico.

Este artigo está estruturado como segue: a Seção 8.1 apresenta a GKB. A Seção 8.2 descreve algumas das aplicações que utilizam as geo-ontologias geradas pela GKB. A Seção 8.3 introduz os resultados obtidos com experimentos para dimensionar a geograficidade[1] de textos em português. A Seção 8.4 descreve o sistema de extração, anotação e integração

[1] Por geograficidade entende-se a quantidade de informação geográfica presente em textos.

de conhecimento geográfico (SEI-Geo), e apresenta as avaliações realizadas com o SEI-Geo. A Seção 8.5 conclui o artigo.

8.1 Geographic Knowledge Base - GKB

A GKB é um dos componentes desenvolvidos no Pólo XLDB da Linguateca (`xldb.di.fc.ul.pt`) em colaboração com o projeto *Geographic Reasoning for Search Engines* (GREASE) (`xldb.di.fc.ul.pt/wiki/grease`), o qual pesquisa métodos, algoritmos e arquiteturas de software para atribuir âmbitos geográficos para recursos da rede e para recolher documentos usando entidades geográficas.

A GKB é um ambiente de extração e integração de conhecimento geográfico que contém informações provenientes de fontes de dados administrativas semi-estruturadas de autoridades junto com um conjunto de regras para integração de informação. A expansão do conhecimento contido na GKB ocorre com informação proveniente de textos. Esses textos são a entrada de informação para o Sistema de Extração, Anotação e Integração de Conhecimento Geográfico (SEI-Geo), que é o responsável por gerar uma representação estruturada do conhecimento geográfico extraído e integrá-lo no repositório da GKB.

A GKB suporta a definição de relacionamentos ontológicos entre entidades, tais como meronímia, sinonímia e adjacência, entre outros. A GKB também suporta relacionamentos inter-domínios, os quais são associações entre entidades de domínios diferentes. Por exemplo, o âmbito geográfico[2] de uma entidade do domínio de rede é representado como um relacionamento entre um sítio da rede (entidade do domínio da Internet) e uma região geográfica (uma entidade do domínio geográfico).

A informação armazenada no repositório da GKB pode ser exportada com uma ferramenta nomeada GOG (*Geographic Ontology Generator*). A GOG permite selecionar partes da informação armazenada na GKB, uma vez que os repositórios da GKB têm, atualmente, cerca de meio milhão de entidades e o usuário raramente quer receber toda a informação. A GOG exporta a informação no formato OWL (`www.w3.org/TR/owl-features/`), uma representação que estende o RDF (`www.w3.org/TR/REC-rdf-syntax/`) e, consequentemente, é também um formato XML. A geo-ontologia completa de Portugal (Geo-Net-PT01) contém mais de 400.000 entidades e é um recurso público disponível em `xldb.fc.ul.pt/geonetpt/`.

[2] Nesse artigo, entende-se âmbito geográfico como a região geográfica, se ela existe, onde a média das pessoas pensa ser mais relevante para uma página, sítio ou domínio da rede. Por exemplo, o âmbito geográfico do sítio da Câmara de Lisboa (`www.cm-lisboa.pt`) é o concelho de Lisboa.

8.2 Aplicações que utilizam as geo-ontologias geradas a partir da GKB

As geo-ontologias exportadas pela GKB têm sido utilizadas por diversas aplicações que incluem: sistemas para reconhecimento de entidades mencionadas (REM), um classificador de documentos de acordo com seu âmbito geográfico, uma interface de recolha de informação para consultas geográficas e uma interface XML para consultas a almanaques geo-temporais, entre outras.

8.2.1 Sistemas de REM

CaGE: é um sistema de REM e de atribuição de âmbito geográfico a páginas da rede (Silva et al., 2006; Martins et al., 2007b). O CaGE utiliza as geo-ontologias geradas a partir da GKB nas fases de identificação e desambiguação de locais (Cardoso et al., 2006b). Martins et al. (2007b) apresentam a arquitetura do CaGE, bem como a descrição detalhada do uso das geo-ontologias.

Faísca: é um sistema de reconhecimento de locais que faz uso dos conceitos e ocorrências contidos nas geo-ontologias geradas a partir da GKB (Cardoso et al., 2008a). O Faísca não explora os relacionamentos existentes entre conceitos nas ontologias, mas utiliza os conceitos para desambiguar nomes de locais.

8.2.2 Módulos de um sistema de recolha de informação geográfica

As geo-ontologias geradas a partir da GKB têm sido utilizadas por diversos módulos do sistema de recolha de informação geográfica da Universidade de Lisboa no GeoCLEF 2007 (Cardoso et al., 2008a).

QueOnde: é um módulo que utiliza as geo-ontologias para dividir o tópico de uma consulta em três partes: 'O que', 'Relacionamento espacial' e 'Onde'. Por exemplo, para o tópico 'tráfego marítimo nas ilhas portuguesas', QueOnde consulta a geo-ontologia e verifica que 'portuguesas' é um adjetivo relativo a Portugal e que 'ilhas' é um conceito geográfico.

QuerCol: é um módulo que utiliza a Geo-Net-PT01 para fazer expansão de consulta. O QuerCol interpreta uma consulta como duas partes: 'O quê' e 'Onde'. A geo-ontologia é usada para expandir o(s) termo(s) da parte 'Onde'. Por exemplo, na consulta 'regiões vinícolas em Portugal', o módulo QuerCol expande o nome Portugal para todas as províncias, distritos, concelhos e freguesias existentes na Geo-Net-PT01 e que fazem parte de Portugal.

Outro módulo do sistema que utilizou as geo-ontologias geográficas é o sistema de reconhecimento de locais Faísca, descrito na seção anterior.

Figura 8.2: Exemplo de interface para recolha de informação geográfica usando a GKB.

8.2.3 Interface de Motor de Pesquisa Geográfica

A GKB é usada também na interface do protótipo Geotumba (`local.tumba.pt`), um sistema para recolha de informação geográfica (ver figura 8.2).

No campo `Local?` o usuário digita a região, a rua, o código postal ou outra entidade geográfica para reduzir o âmbito da consulta. Quando um nome geográfico ambíguo é detectado na consulta, Geotumba apresenta as possíveis alternativas para desambiguação da mesma. Por exemplo, o nome "rua Castelo Branco" ocorre em cinco concelhos diferentes na Geo-Net-PT01, os quais são apresentados no lado esquerdo da figura 8.2. Além da consulta por texto, o usuário pode utilizar os mapas para definir o âmbito da consulta.

8.2.4 Interface para consultas a almanaques geo-temporais

A Geo-Net-PT01 também é utilizada no projeto DIGMAP (*Discovering our Past World with Digitised Maps*, `www.digmap.eu`) (Borbinha et al.), especificamente em uma interface XML para consultas a almanaques geo-temporais. Neste serviço, a Geo-Net-PT01 é integrada com outros almanaques existentes considerando a dimensão temporal juntamente com o conteúdo geográfico dos almanaques. A figura 8.3 apresenta a interface do sistema.

Para cada local inserido pelo usuário, o sistema de consultas a almanaques geo-temporais percorre os almanaques e apresenta o nome do local juntamente com seus metadados, relacionamentos e população, entre outras informações subjacentes a cada almanaque. A informação geográfica é apresentada em diversas linguagens (e.g. XML, OWL e KML - *Keyhole Markup Language*, `www.opengeospatial.org/standards/kml/`), conforme o almanaque as disponibiliza.

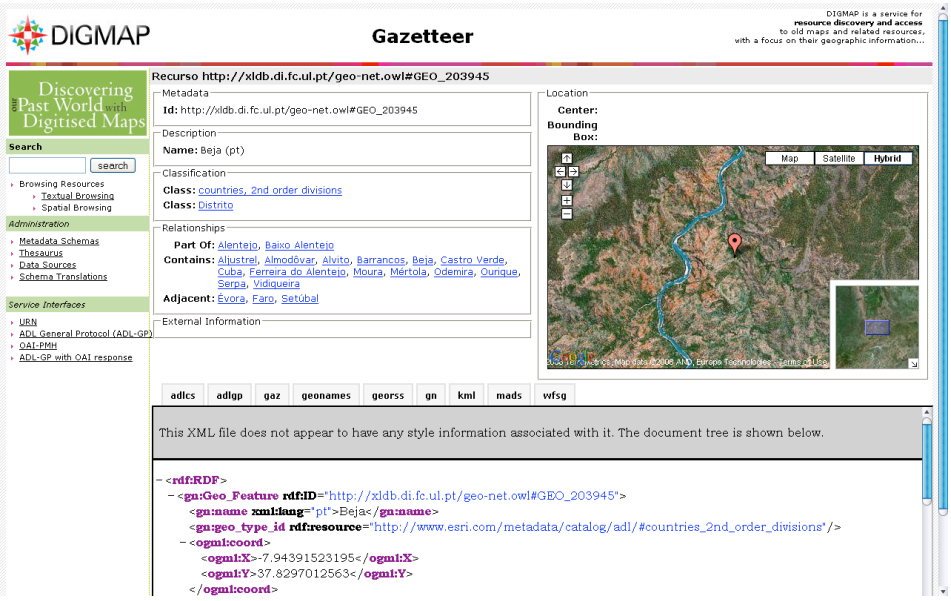

Figura 8.3: Interface para Consultas a Almanaques Geo-temporais.

No exemplo da figura 8.3, o sistema apresenta os metadados sobre o 'distrito de Beja', os quais incluem os relacionamentos de `parte-de`, `contém` e `adjacência`. Na parte inferior da figura, estão nove almanaques que contêm informação sobre o 'distrito de Beja'. No canto superior direito, o 'distrito de Beja' é ilustrado no mapa.

A Geo-Net-PT01 já foi requisitada por dezenas de investigadores, na sua maioria de Portugal e do Brasil, evidenciando o interesse da comunidade em estruturas de representação de conhecimento geográfico. A figura 8.4 apresenta a distribuição geográfica dos pedidos por países.

Por fim todo o conteúdo das geo-ontologias geradas pela GKB pode ser visualizado com a interface Geobase, apresentada na figura 8.5 (`www.tumba.pt/tumba/geobase`).

As aplicações que utilizam as geo-ontologias geradas pela GKB necessitam de informação geográfica além daquela proveniente de fontes de informação estruturadas e semi-estruturadas. Nomes históricos e alternativos de locais, por exemplo, ainda não estão na GKB, mas podem ser encontrados em textos. Programas foram implementados para dimensionar a geograficidade de textos em português e para conhecer a sobreposição da informação armazenada na GKB com a informação geográfica em textos.

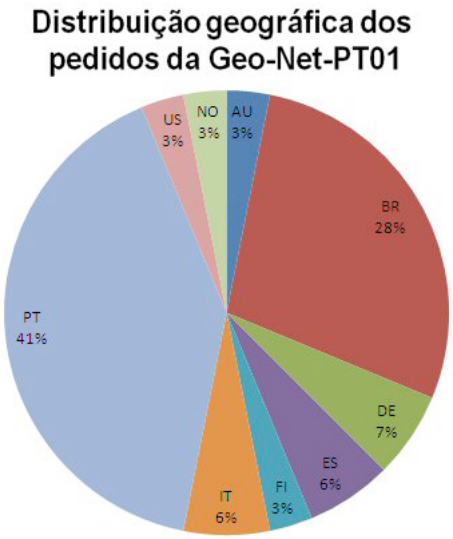

Figura 8.4: Distribuição geográfica dos pedidos da Geo-Net-PT01 por países.

Figura 8.5: Geobase: interface de visualização da Geo-Net-PT01.

8.3 Geograficidade de textos

Para verificar a geograficidade em textos da rede, foi utilizado o WPT 03, uma recolha da rede portuguesa de 2003, com 12 GB, 3.7 milhões de páginas e 1.6 bilhões de palavras (`www.linguateca.pt/WPT03`) (Cardoso et al., 2007). Aproximadamente 68.6% dessas páginas estão em português e mais de 1.5 milhões são distintas. O sistema de REM utilizado foi o SIEMÊS (Sarmento, 2006), que na avaliação conjunta do Primeiro HAREM (Santos et al., 2006) alcançou 70% de precisão e 75% de abrangência para a categoria local. Entretanto, a versão utilizada em nossos experimentos é uma versão com melhoramentos sobre aquela utilizada no Primeiro HAREM.

A caracterização da informação geográfica em textos passa pela verificação da presença de nomes de locais em nomes de pessoas e organizações. Considerando uma amostra aleatória de 32.000 documentos da rede, os resultados evidenciam que 31% das entidades mencionadas distintas da categoria pessoa e 23,43% das entidades mencionadas distintas da categoria organização contêm um nome geográfico incluído na Geo-Net-PT01.

Para investigar se o tipo de local ocorrendo em textos da rede portuguesa tinha diferentes propriedades (granularidade, geografia física (rios, montanhas, etc.)), foram verificados os tipos das entidades mencionadas da categoria local que o SIEMÊS encontrou após ser executado sobre a mesma amostra de 32.000 documentos. O resultado mostrou que 85% dos tipos de locais reconhecidos pelo SIEMÊS estão concentrados em apenas três (povoamento, endereço completo e sociedade/cultura) dos tipos de locais definidos no Primeiro HAREM. Estatísticas mais detalhadas sobre a geograficidade em textos podem ser consutadas em Chaves e Santos (2006).

Quanto aos tipos de arruamentos, os predominantes na geografia administrativa de Portugal são ruas e travessas. Somente ruas representam mais de 60% dos tipos de arruamentos do país. Rua também é o tipo de arruamento mais freqüente no WPT 03, após o tipo ambíguo acesso. Por outro lado, as travessas ocorrem com bem menos freqüência no WPT 03, sendo apenas o 28º tipo de arruamento mais freqüente.

Cera de 60% dos nomes[3] presentes na Geo-Net-PT01 estão presentes no WPT 03. Aqueles compostos por quatro palavras são os menos freqüentes, ao passo que os nomes formados por uma palavra atingem quase 80% de presença nesse corpus da rede. Outros resultados sobre a presença de informação geográfica de ontologias em textos e sobre a ambigüidade existente entre nomes de uma ontologia são descritos em (Santos e Chaves, 2006).

Após verificar que existe informação geográfica em textos suficiente para expandir geo-ontologias, foi desenvolvido o SEI-Geo.

[3] Neste contexto, todos os nomes na Geo-Net-PT01 exceto nomes de arruamentos e códigos-postais.

8.4 Sistema de Extração, Anotação e Integração de conhecimento Geográfico - SEI-Geo

O SEI-Geo foi desenvolvido no Pólo XLDB da Linguateca no âmbito do projeto GREASE e tem como objetivo reconhecer o conhecimento geográfico disponível em textos, gerar uma representação estruturada desse conhecimento e integrá-lo em geo-ontologias. O sistema é composto por dois módulos principais: o de Extração de Informação Geográfica (EIG) e o de Integração de Conhecimento Geográfico (ICG).

O EIG recebe como entrada um conjunto de textos que são segmentados em frases. O EIG contém uma quantidade abrangente de regras que indicam a presença de conceitos e relacionamentos nas frases. Tais frases, juntamente com conceitos de geo-ontologias, são a entrada de uma função que extrai frases com potencial conteúdo geográfico. Essas frases são a entrada de dois sub-módulos: o extrator de arbustos[4] e o anotador. O extrator de arbustos detecta ocorrências geográficas e relacionamentos semânticos e tem uma função de filtro, na qual o conteúdo geográfico, duplicado ou sobreposto, é eliminado. O resultado desse processo é um conjunto de arbustos que são utilizados como entrada no ICG. O anotador insere etiquetas com nomes de categoria semântica, tipo e subtipo. O anotador também possui a capacidade de reconhecer e anotar relacionamentos entre locais.

O ICG recebe os arbustos extraídos e anotados, e o conhecimento armazenado na GKB, faz a integração e retorna para a GKB o conhecimento geográfico expandido. A integração de conhecimento geográfico em geo-ontologias concentra-se em encontrar informação geográfica complementar àquela existente nas geo-ontologias e integrar essa informação no nível de granularidade mais adequado nas geo-ontologias. A integração de conhecimento geográfico com o SEI-Geo ocorre quando novos fatos geográficos são descobertos em texto.

8.4.1 Avaliação do SEI-Geo

O SEI-Geo tem sido avaliado na sua capacidade de extrair, anotar e integrar conhecimento geográfico. O SEI-Geo participou no Segundo HAREM (Mota e Santos, 2008) e conseguiu atingir resultados satisfatórios no cenário seletivo de identificação e classificação de locais. Considerando somente a medida F, o SEI-Geo foi o segundo melhor sistema nesse cenário com 0,5953, enquanto o melhor sistema atingiu 0,6246 na tarefa de classificação semântica. A participação do SEI-Geo no Segundo HAREM é descrita em Chaves (2008).

Além da tarefa de anotação de textos, o SEI-Geo foi avaliado, através de testes de mutilação, na sua capacidade de extrair locais e recompor uma geo-ontologia existente. Testes de mutilação consistem na destruição de parte de um objeto de estudo e na sua reconstrução. Especificamente quando se trata de estruturas de representação de conhecimento

[4] Um *arbusto* é composto por pelo menos duas entidades geográficas candidatas a locais e um relacionamento. Esse conjunto forma uma tripla. Não há número máximo de entidades e relacionamentos pré-definido.

	Público 1994	Público 1995	FSP 1994	FSP 1995
SEI-Geo reconstruido	148 (70,47%)	161 (76,30%)	117 (62,56%)	109 (60,55%)
ISO-3166-1 na coleção	210	211	187	180

Tabela 8.1: Resultado do teste de mutilação para países e territórios nos corpora jornalísticos.

tal como ontologias, um (ou vários) nível da hierarquia de conceitos e ocorrências é destruído e a partir de informação textual tenta-se reconstruir a informação retirada inicialmente. Para implementar esse teste, foram retiradas todas as ocorrências do tipo de entidade `ISO-3166-1` (que corresponde a países e territórios) da ontologia WGO (Martins et al., 2007a). Todos os arbustos extraídos pelo SEI-Geo que contêm o tipo de entidade `ISO-3166-1` foram enviados à geo-ontologia com o objetivo de encontrar um identificador para cada entidade geográfica reconhecida. A tabela 8.1 apresenta os resultados dos testes de mutilação usando o corpus CHAVE (Santos e Rocha, 2005).

Conforme a tabela 8.1, o Público é uma fonte mais rica em informação geográfica ao nível de países e territórios do que o Folha de São Paulo. Dos 211 países e territórios existentes na parte do Público relativa ao ano de 1995, 161 (70,30%) foram reconhecidos e representados em triplas no formato de arbusto. Das 238 ocorrências do tipo de entidade `ISO-3166-1` da WGO, 211 ocorrem nesse corpus. Um dos fatores que levam o Público a conter mais locais da WGO é os nomes de locais estarem na sua maioria descritos no português de Portugal. Exemplos desses casos encontrados no Público e ausentes no Folha de São Paulo são: 'Coreia do Sul', 'Eslovénia' e 'Ilhas Caimão'. Os resultados dos testes de mutilação indicam que o SEI-Geo é capaz reconstituir uma geo-ontologia recebendo como entrada conceitos sem ocorrências.

Quanto à expansão de geo-ontologias, o SEI-Geo recebe como entrada um corpus e geo-ontologias e devolve como resultado um conjunto de arbustos com as geo-ontologias enriquecidas com novos locais e relacionamentos reconhecidos no corpus. Se o SEI-Geo encontra uma ocorrência de um conceito e essa ocorrência já está na geo-ontologia, o resultado permite validar a ocorrência e a geo-ontologia não é expandida.

A primeira avaliação foi realizada com a parte do corpus CHAVE relativa ao ano de 1995. De um total de 50.495 arbustos, foi selecionada aleatoriamente uma amostra de 100 arbustos compostos por 143 triplas. Cada tripla dessa amostra foi avaliada manualmente de acordo com os seguintes critérios:

Integrável (I): quando as duas entidades geográficas da tripla forem realmente locais e a relação entre elas estiver correta.

Integrável com Assistência (IA): quando duas entidades geográficas forem corretas e não existir relacionamento explícito no texto ou o algoritmo não conseguiu identificar. Nesse caso o avaliador deve inserir o relacionamento correto.

Existente (E): quando as entidades geográficas e o relacionamento reconhecido entre essas entidades geográficas já está em pelo menos uma das ontologias.

Falso (F): quando no máximo uma entidade geográfica da tripla é um local ou as duas entidades geográficas não possuem relacionamento no mundo real.

Após a avaliação das 143 triplas, eu encontrei 2 I, 61 IA, 19 E e 61 F. Esses resultados indicam que a maior parte das triplas integráveis são integráveis com assistência. Ainda resta um número elevado de triplas falsas, mas esses valores já eram esperados dados os resultados da participação do SEI-Geo no Segundo HAREM.

8.5 Considerações Finais

Este artigo resumiu meu trabalho no âmbito da Linguateca ao longo dos últimos anos. A base de conhecimento geográfico armazena o conteúdo exportado como geo-ontologias que estão disponíveis publicamente. Esse conteúdo geográfico é expandido com informação textual extraída pelo SEI-Geo. O SEI-Geo foi avaliado no Segundo HAREM no que diz respeito à sua capacidade de anotação de locais e também apresentou resultados encorajadores nos testes de mutilação e expansão de geo-ontologias.

Após as geo-ontologias terem sido utilizadas por várias aplicações, torna-se essencial a criação de uma geo-ontologia mundial com nomes de locais em português, abrangendo as variantes da língua de Portugal e do Brasil. Essa nova geo-ontologia pode ser criada reutilizando o modelo-base no qual a GKB foi concebida.