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

Marcirio Silveira Chaves
Pólo XLDB da Linguateca
LaSIGE - Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa

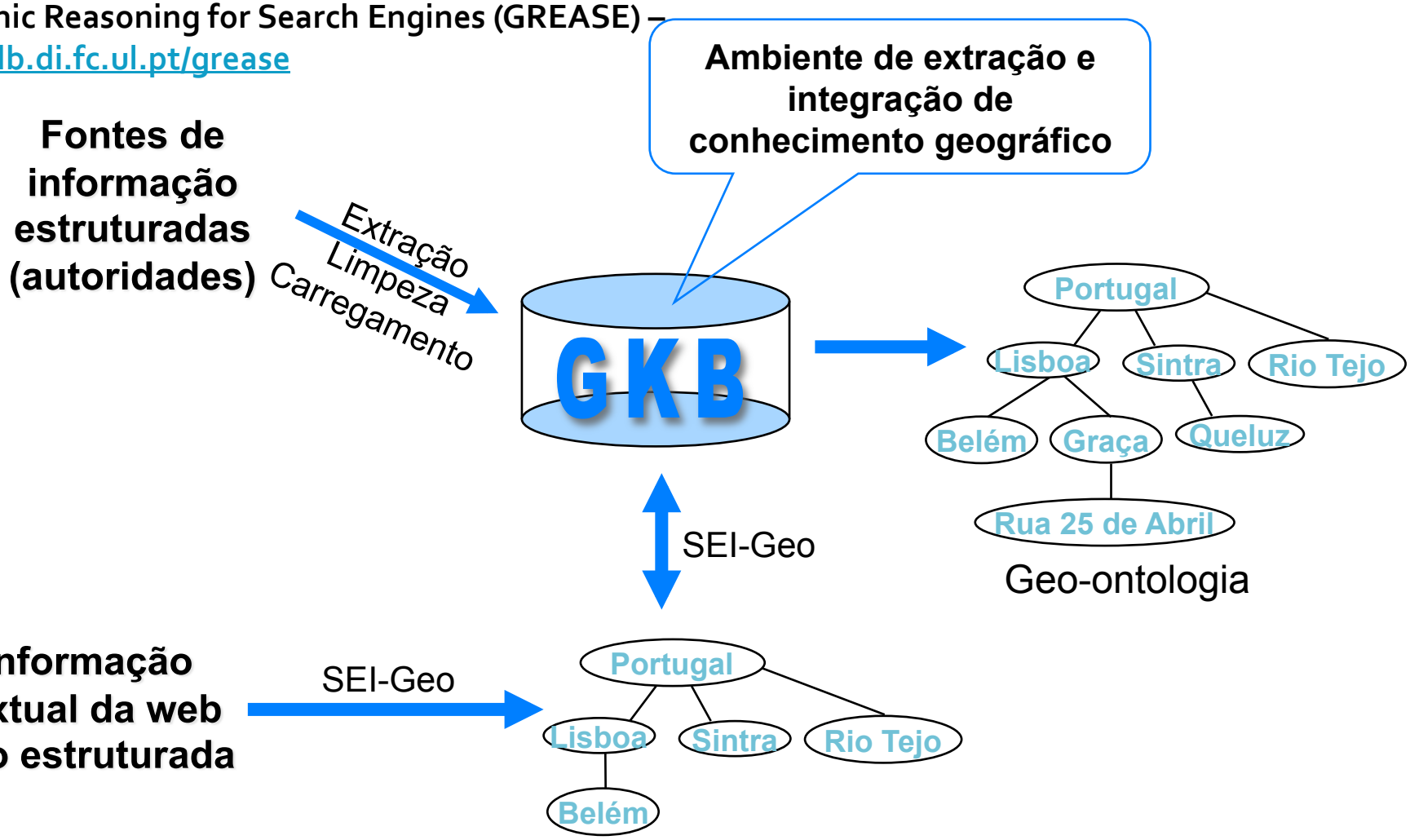
Encontro 10 Anos da Linguateca
PROPOR 2008 - Aveiro - Portugal

Enquadramento

- Web Semântica
- Disponibilização de recursos
  - Geo-ontologias
- Sistema de Extração, anotação e Integração de Conhecimento Geográfico – SEI-Geo

Geographic Knowledge Base -GKB

Geographic Reasoning for Search Engines (GREASE) – http://xldb.di.fc.ul.pt/grease



Geo-ontologias

- **Geo-Net-PT**
  - + de 400k entidades geográficas
  - Todo o vocabulário do domínio geográfico administrativo de Portugal
  - Recurso público e gratuito: http://xldb.fc.ul.pt/geonetpt
- **WGO (*World Geographic Ontology*)**
  - nomes, conceitos e relacionamentos das principais divisões administrativas do mundo
    - países e territórios até cidades com **mais de 100k habitantes**,
    - entidades geográficas no domínio físico, tais como oceanos, mares e montanhas
  - + de 13k entidades geográficas
  - + de 24k relacionamentos (parte-de e adjacência)

Aplicações que Utilizam as Geo-ontologias Geradas a partir da GKB

- Sistemas de REM
  - **CAGE**
    - sistema de **REM** e de atribuição de **âmbito geográfico** a páginas da rede.
    - usa **geo-ontologias** geradas a partir da GKB nas fases de **identificação** e **desambiguação de locais**
  - **Faísca**
    - sistema de **reconhecimento de locais** que usa conceitos e ocorrências das **geo-ontologias**
    - não explora os relacionamentos existentes entre conceitos nas geo-ontologias
    - utiliza os **conceitos** para **desambiguar** nomes de locais na fase de REM

Aplicações que Utilizam as Geo-ontologias Geradas a partir da GKB

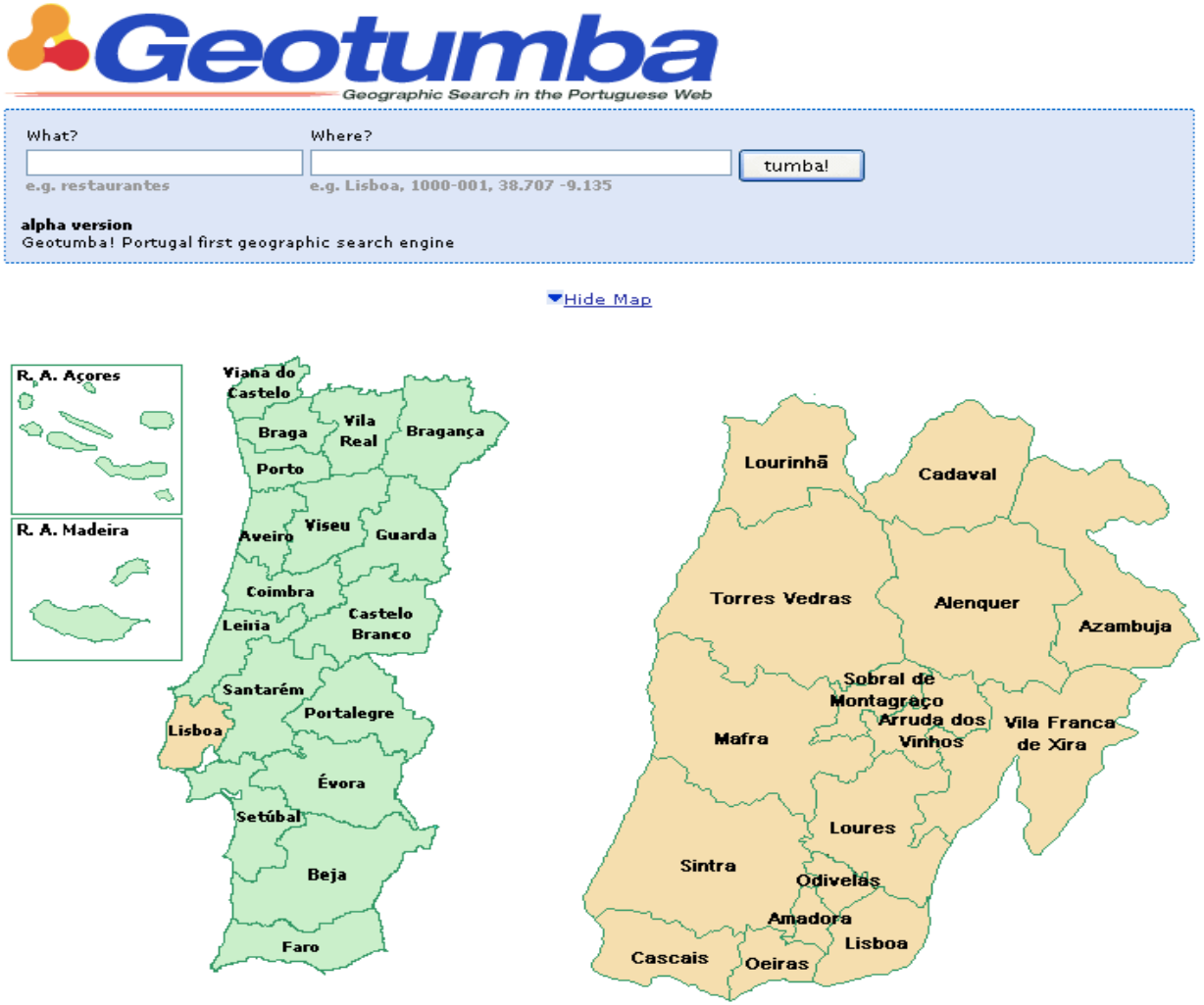
- Módulos de um Sistema de Recolha de Informação Geográfica
  - **QueOnde**
    - utiliza as geo-ontologias para dividir o tópico de uma consulta em três partes: `**O que**', `**Relacionamento espacial**' e `**Onde**'
    - Tópico: `tráfego marítimo nas ilhas portuguesas'
      - `portuguesas' é um ***adjetivo*** de Portugal
      - `ilhas' é um **conceito** geográfico

Aplicações que Utilizam as Geo-ontologias Geradas a partir da GKB

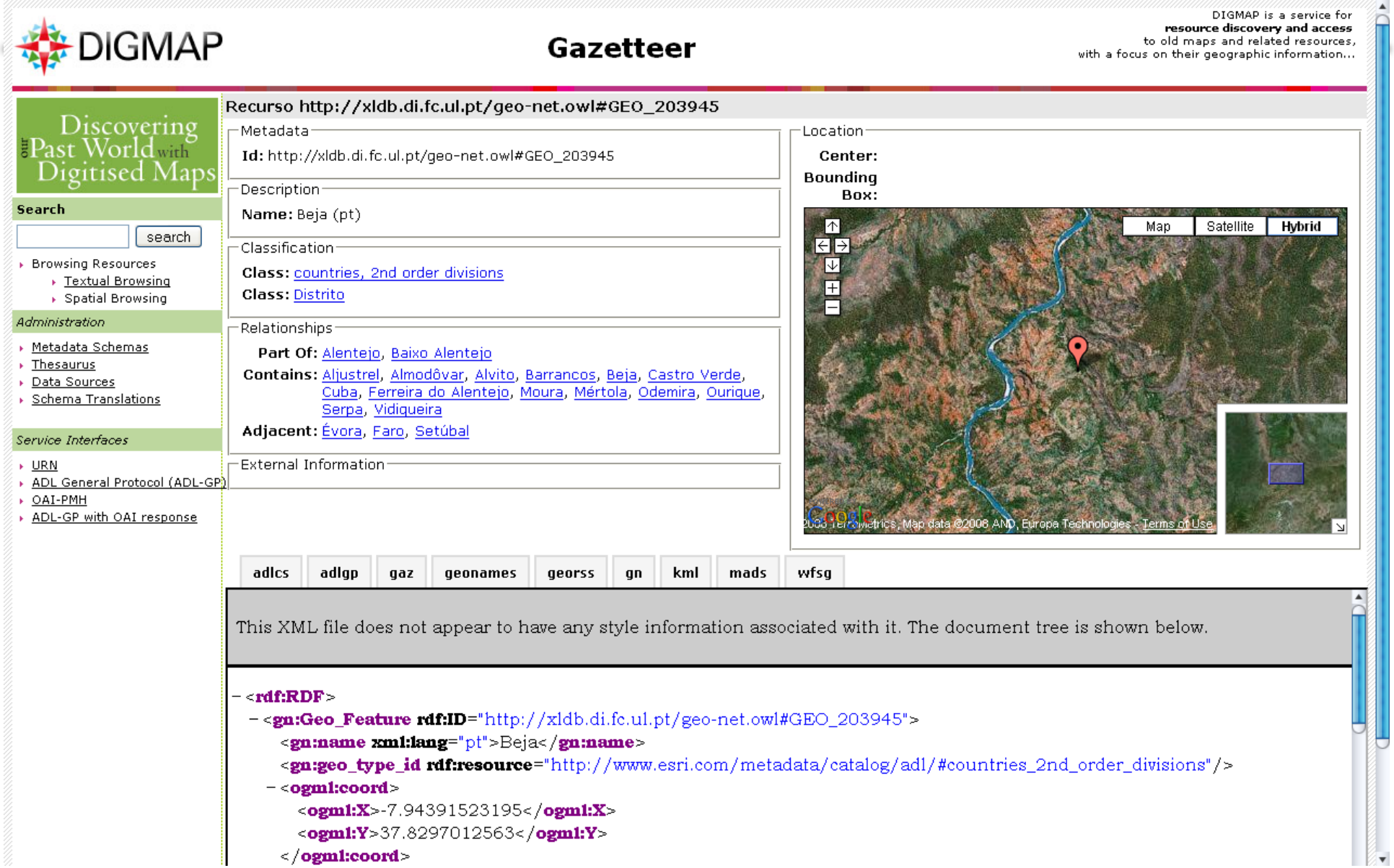
- Módulos de um Sistema de Recolha de Informação Geográfica
  - **QuerCol**
    - usa geo-ontologia para expansão de consulta
    - interpreta consulta como duas partes: `**O que**' e `**Onde**'
    - geo-ontologia: expandir o(s) termo(s) da parte `Onde'
    - consulta: `regiões vinícolas em Portugal'
      - expande o nome Portugal para todas as províncias, distritos, concelhos e freguesias da geo-ontologia e que fazem parte de Portugal

Aplicações que Utilizam as Geo-ontologias Geradas a partir da GKB

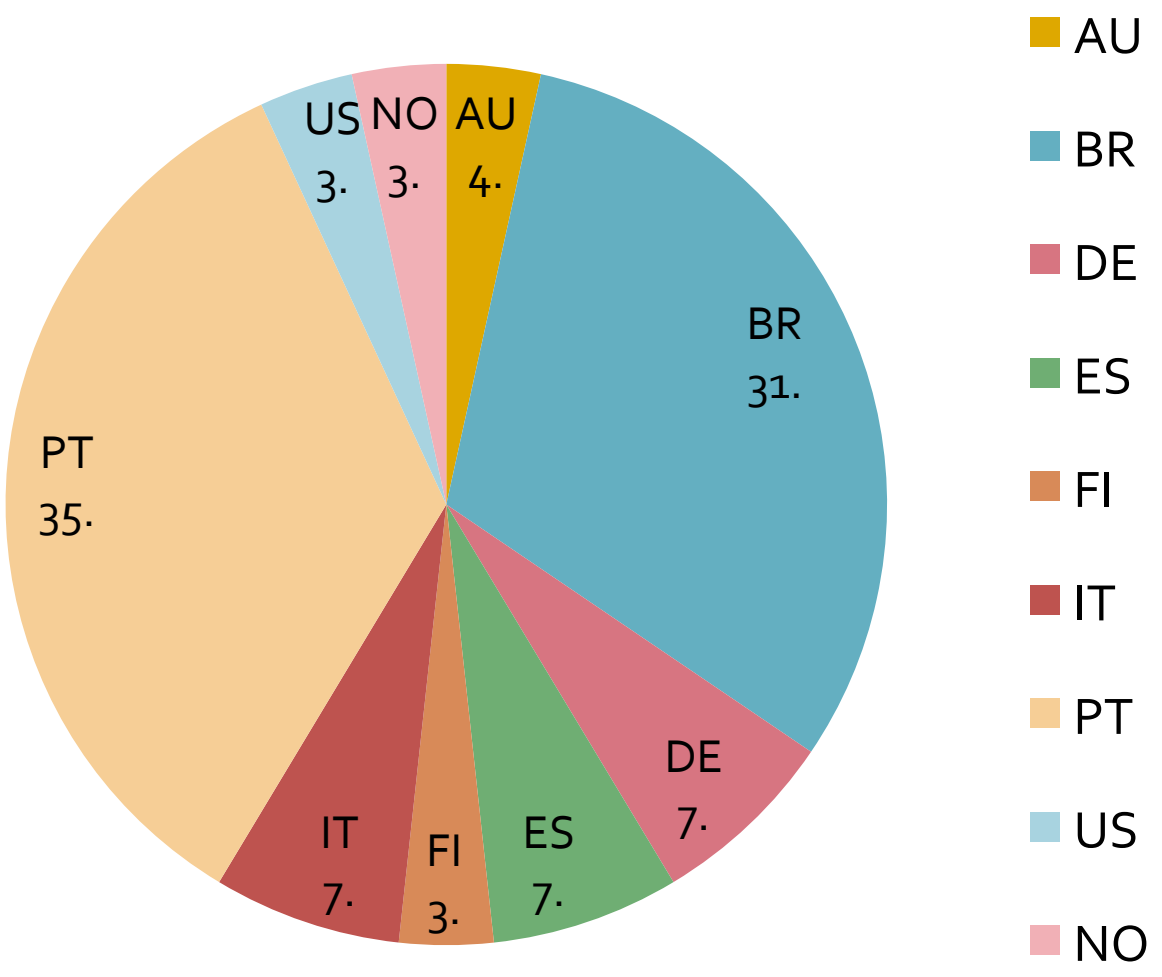
- Interface do Sistema de RIG



Interface para Consultas a Almanaques Geo-temporais



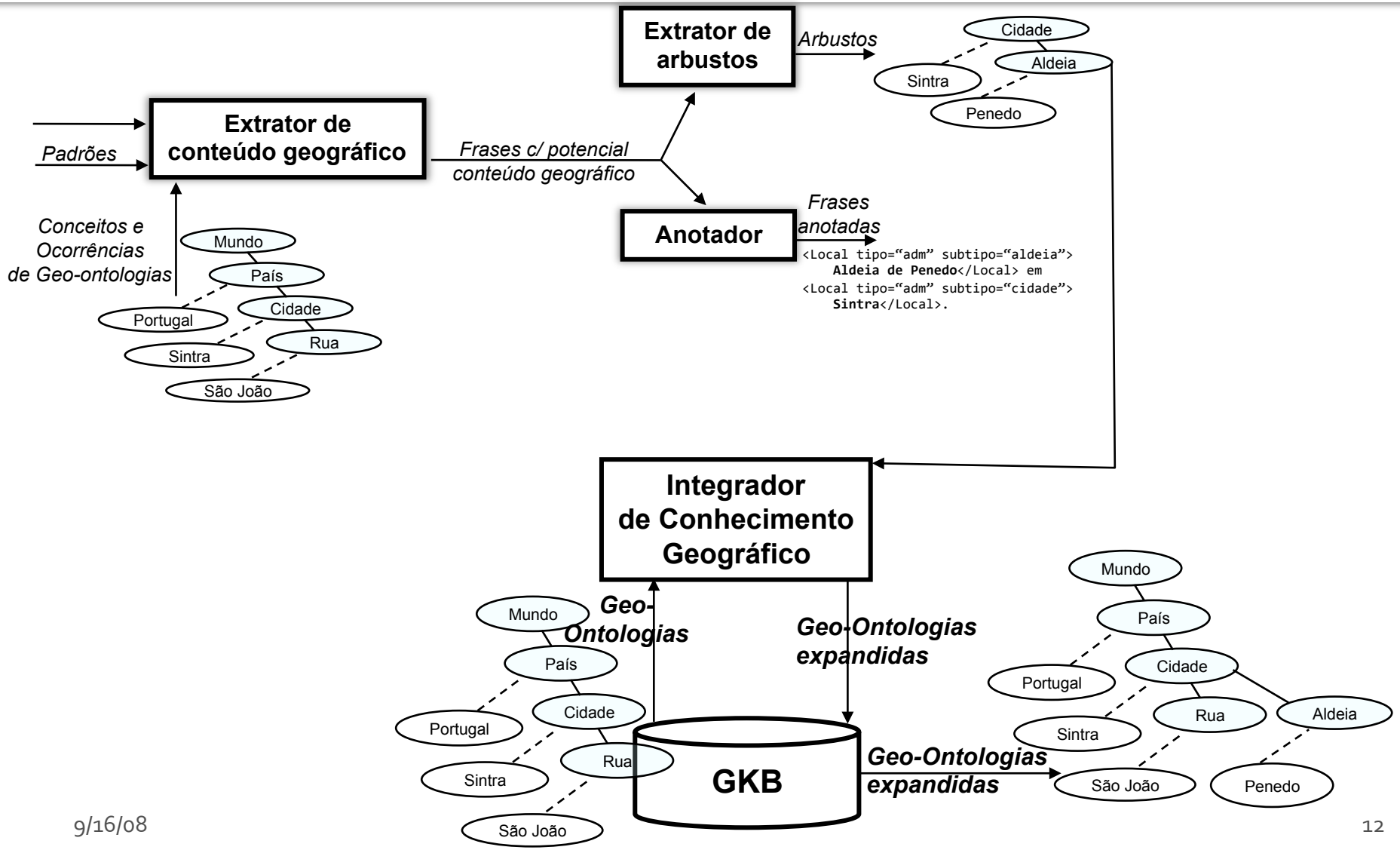
Distribuição geográfica dos pedidos da Geo-Net-PT por países



Geograficidade de Textos

- WPT03 - http://linguateca.di.fc.ul.pt/WPT03
- Caracterização da informação geográfica em textos
  - nomes de locais em nomes de pessoas e organizações
  - amostra aleatória de 32.000 documentos do WPT03
    - **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-PT**

Sistema de Extração, anotação e Integração de conhecimento Geográfico - SEI-Geo



Sistema de Extração, anotação e Integração de conhecimento Geográfico - SEI-Geo

- Avaliação
  - Extração e Anotação – Eventos HAREM
  - Integração – Testes de Mutilação e Expansão de Geo-ontologias

- Testes de Mutilação
  - Resultado do teste de mutilação para países e territórios nos corpora jornalísticos.

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

Sistema de Extração, anotação e Integração de conhecimento Geográfico - SEI-Geo

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

- Público
  - nomes de locais na sua maioria -> português de Portugal
  - Exemplos de casos encontrados no Público e ausentes na Folha de São Paulo
    - `Coreia do Sul', `Eslovénia' e `Ilhas Caimão'
- Resultados dos testes de mutilação indicam que o SEI-Geo é capaz reconstituir uma geo-ontologia recebendo como entrada conceitos sem ocorrências

Considerações Finais

- Resumo do meu trabalho no âmbito da Linguateca ao longo dos últimos anos
  - metodologia para construção e expansão de geo-ontologias
  - GKB - conteúdo exportado como geo-ontologias disponíveis publicamente
  - avaliação do SEI-Geo
    - Segundo HAREM
    - testes de mutilação
- Trabalho Futuro
  - criação de uma geo-ontologia mundial
    - nomes de locais em português: variantes da língua de Portugal e do Brasil

Referências

- Chaves, Marcirio Silveira. **Geo-ontologias e padrões para reconhecimento de locais em textos: a participação do SEI-Geo no Segundo HAREM**. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM, 2008. (a aparecer)

- Manguinhas, H.; Martins, B. & Borbinha, J. **Geo-Temporal Web Gazetteer Service Integrating Data From Multiple Sources.** 3rd IEEE International Conference on Digital Information Management, IEEE, London, UK, University of East London, November 2008. (a aparecer)

- Martins, Bruno; Silva, Mário J.; Chaves, Marcirio Silveira. **O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa.** In: Diana Santos & Nuno Cardoso (eds.). Documentação e actas do HAREM, a primeira avaliação conjunta na área. p. 97-112, 2007.

- Santos, Diana e Chaves, Marcirio Silveira. **The place of place in geographical IR**. In 3rd Workshop on Geographic Information Retrieval, SIGIR'2006. pages 5-8, August 10th, Seattle, 2006.

- Chaves, Marcirio Silveira e Santos, Diana. **What kinds of geographical information are there in the Portuguese Web?**. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno Mamede, Claudia Oliveira & Maria Carmelita Dias (eds.), Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'2006), LNAI 3960 - Springer, (Itatiaia, RJ, 13 a 17 de maio), pp. 264-267.

- Martins, Bruno; Cardoso, Nuno; Chaves, Marcirio Silveira; Andrade, Leonardo; Mário J. Silva. **The University of Lisbon at GeoCLEF 2006**. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, CLEF, volume 4730 of Lecture Notes in Computer Science, pages 986-994. Springer, 2006.

- Chaves, Marcirio Silveira; Silva, Mário J. and Martins, Bruno. **A Geographic Knowledge Base for Semantic Web Applications**. 20th Brazilian Symposium on Databases - SBBD, Uberlândia, Minas Gerais, Brazil, pp. 40-54, 3-7 October, 2005.