Geographic IR Challenges

by Nuno Cardoso

Faculty of Sciences,
University of Lisbon, LASIGE

Presentation held at SINTEF ICT, Oslo, Norway, 4th December, 2007

LaSIGE Laboratory

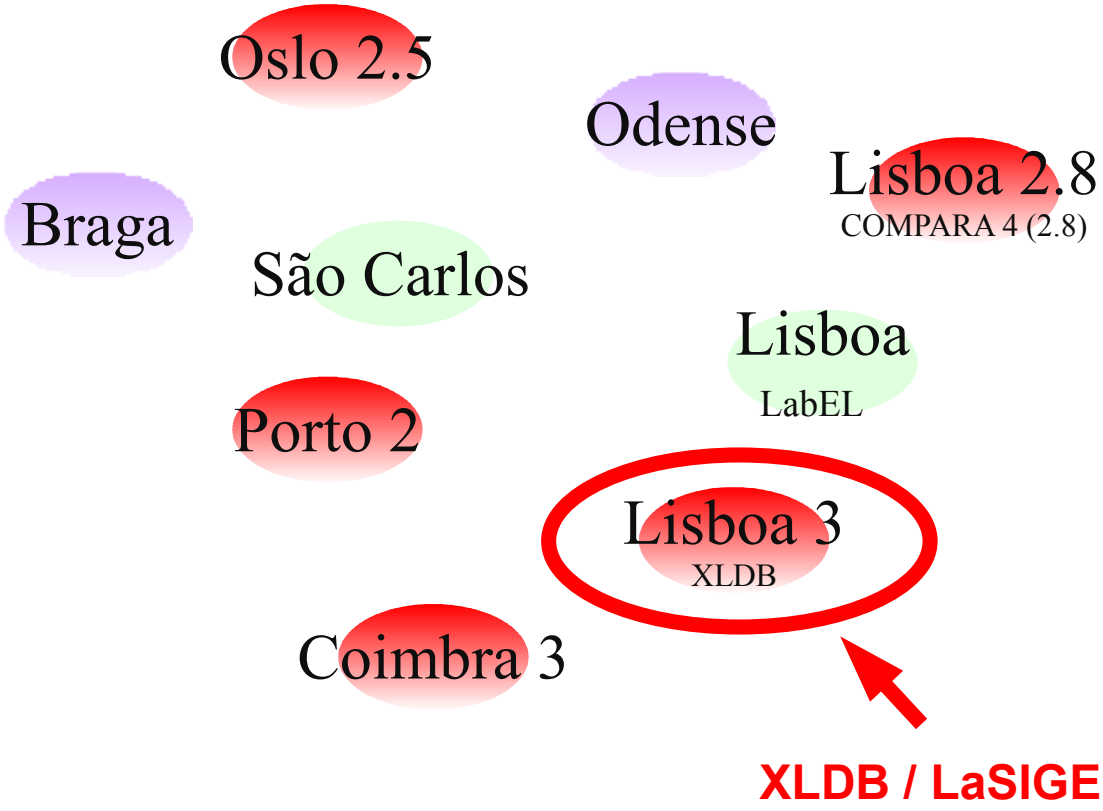
- Large-Scale Informatics System Laboratory
- 4 research groups: **XLDB**, Navigators, DIALNP, HCIM
- XLDB research lines: Information retrieval, text mining, natural language processing, web archiving and search, information visualization, bioinformatics
- XLDB's staff: Ana Paula Afonso, Ana Vaz, André Falcão, Catarina Rodrigues, Cátia Pesquita, Daniel Faria, Daniel Gomes, **David Cruz**, Francisco Couto, Hugo Bastos, Leonardo Andrade, Liliana Moreira, Luís Russo, **Marcirio Chaves**, **Mário Silva**, **Nuno Cardoso**, Paulo Carreira, Paulo Pombinho, Sérgio Freitas, Tiago Grego

(In bold: Linguateca 's XLDB Node)

Linguateca, a project for Portuguese

- A distributed resource center for Portuguese language technology

- IRE model:
  - **I**nformation
  - **R**esources
  - **E**valuation



www.linguateca.pt

Human capital of Linguateca
(in bold: intersection with XLDB)

Senior researchers: *Diana Santos, José João Almeida, Eckhard Bick, Belinda Maia, Ana Frankenberg Garcia,* ***Mário J. Silva****, Paulo Gomes, Luís Costa*

Researchers: *Luís Miguel Cabral, Susana Inácio, Rosário Silva, Paulo Rocha, Cláudia Freitas (PhD), Sérgio Matos (PhD), Hugo Oliveira, Pedro Martins,* ***David Cruz***

PhD students: ***Marcirio Chaves****, Alberto Simões, Nuno Seco, Anabela Barreiro,* ***Nuno Cardoso***

Former: *Rachel Aires* (PhD, 2005), *Signe Oksefjell*, *Susana Afonso*, *Raquel Marchi*, *Renato Haber, Alex Soares, Pedro Moura, Ana Sofia Pinto, Débora Oliveira, Isabel Marcelino, Cristina Mota, Luís Sarmento, António Silva, Rui Vilela*

In all, 36 (22) persons with some link to Linguateca (not counting trainees or undergraduate students)

Presentation overview

1) What is GIR?
   - GIR terminology
   - Motivation for GIR
2) GIR approaches
   - The case of GeoTumba
3) GIR evaluation
   - In-house evaluation of GIR components
   - International evaluation contests: GeoCLEF
4) My PhD research plan

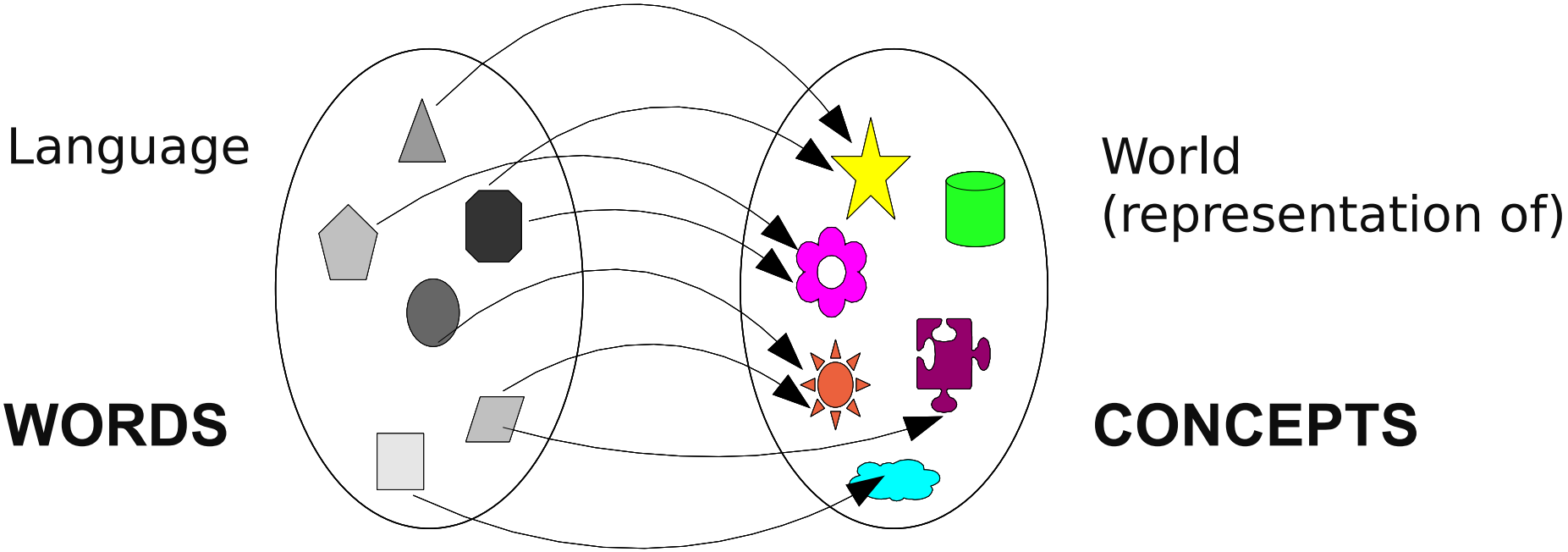
1) What is GIR?

- GIR terminology
- Motivation for GIR

GIR terminology

Why is it important?

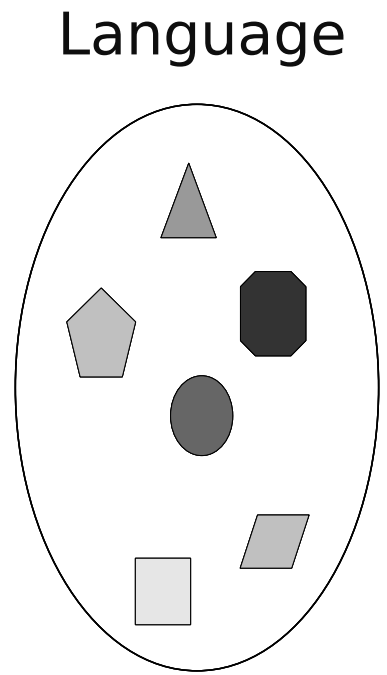
- We are dealing with **words** written in **natural language** that correspond to **concepts**.
- Mapping words in context to concepts is not straightforward...



GIR terminology

Words with geographic meaning:

- **Placenames** – words used to name places. For example: *Oslo*, *Trøndelag*, *Norway, Scandinavia, "Syden", Middle East*.
- **Common nouns** – words denoting classes of places. For example: *island*, *mountain*, *lake*, *city*, *country*, *continent*.
- **"Relators"** – words used to establish a relation between entities (verbs, prepositions, adverbs, etc). For example: "*near*", "*located in*", "*on the shores of*", "*up*".

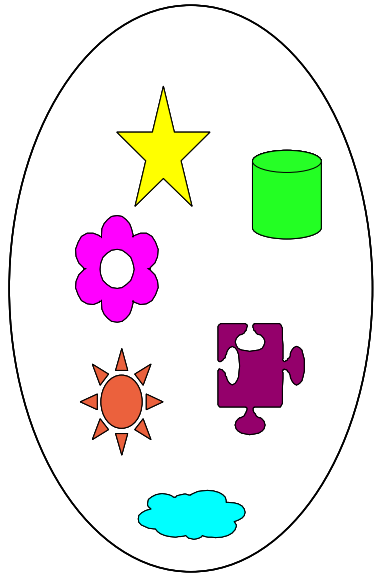


GIR Terminology (cont.)

Geographical concepts:

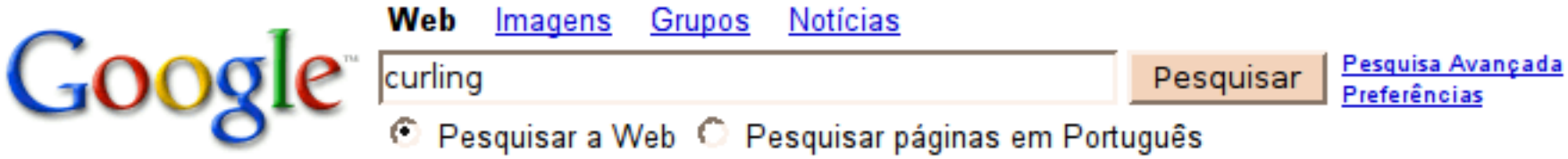
- **Features (ISO-19109)** – an unambiguous location. It can be described by one or more placenames. For example: *Paris*.
- **Feature Types (ISO-19109)** – classes of features. For example, *island*, *mountain*, *lake* (physical), *city*, *continent*, *NUT-3* (administrative). A feature has only one feature type.
- **Relations** – Links joining features **OR** feature types: *part of, adjacent, capital of, etc*. Examples: [Oslo] *part-of* [Norway], [city] *part-of* [country].

World (represen-tation of)



Geographic Information Retrieval

- Geographic Information Retrieval (GIR) = Information Retrieval (IR) + geographic reasoning!
- **Goal**: research methods to retrieve relevant information in a document collection.
- Relevance now comes twofold:
  - **Thematical** (ex: “Find me documents about curling!”)
  - **Geographical** (ex: “in the Scandinavian countries”)



Web Resultados **1** - **10** de cerca de **13.300.000** para **curling**. (**0,28** segundos)

Curling - Wikipedia, the free encyclopedia - [Traduzir esta página]
Curling is a team sport with similarities to bowls and bocce, played on a rectangular sheet of carefully prepared ice by two teams of four players each. **...**
en.wikipedia.org/wiki/**Curling** - 117k - Em cache - Páginas semelhantes

Curling's Premiere Directory - [Traduzir esta página]
This Canadian Premiere **Curling** website details **curling** clubs, **curling** links and championships of this popular winter ice **curling** sport.
www.**curling**.com/ - 108k - Em cache - Páginas semelhantes

World **Curling** Federation - HOME - [Traduzir esta página]
Joomla - the dynamic portal engine and content management system.
www.world**curling**.org/ - 28k - Em cache - Páginas semelhantes

Curling Basics - [Traduzir esta página]
Anhand von Flash-Animationen werden **Curling** - Begriffe und Regeln erklärt. Animated examples of **curling** terms and rules.
www.**curling**basics.com/ - 3k - Em cache - Páginas semelhantes

United States **Curling** Association - USCA **CURLING** HOME - [Traduzir esta página]
The official website for the Olympic sport of **curling** in the United States.
www.usacurl.org/ - 34k - Em cache - Páginas semelhantes

International Olympic Committee - **Curling** - [Traduzir esta página]
Includes history, equipment, events, glossary, and photographs.
www.olympic.org/uk/sports/programme/index_uk.asp?SportCode=CU - 28k -

Google search for “curling”... OK.

Google

Web Imagens Grupos Notícias

curling in scandinavian countries | Pesquisar | Pesquisa Avançada | Preferências

Pesquisar a Web Pesquisar páginas em Português

Web Resultados **1** - **10** de cerca de **48.600** para **curling in scandinavian countries**. (**0,23** segundos)

Four days of Scotland in Sweden - [Traduzir esta página]
"We are also looking to use Sweden's huge interest in **curling** to help enhance our **...** "There are many ties between our two **countries**, especially on the **...**
www.scottishexecutive.gov.uk/News/Releases/2002/09/2288 - 24k -
Em cache - Páginas semelhantes

Curling - [Traduzir esta página]
Curling Basics from The Virtual Library for Sport. **...** Scotland historically have the strongest representation, but the **Scandinavian Countries**, Switzerland, **...**
sportsvl.com/ball/**curling**home.htm - 7k - Em cache - Páginas semelhantes

Scandinavian Heritage Society Organization - [Traduzir esta página]
Members have an opportunity to join the **Curling** League for weekly games. **...** The five **Scandinavian countries** are represented in the Society by local groups: **...**
www.**scandinavian**heritagesociety.org/ - 51k - Em cache - Páginas semelhantes

What is **Curling** How do you Play at Fentons Rink **Curling** in Kent - [Traduz
The major **curling countries** are Scotland, Canada, the **Scandinavian countries**, Switzerland, Germany, USA, Holland, Italy and France. **...**
www.fentonsrink.co.uk/what.html - 7k - Em cache - Páginas semelhantes

The roarin' game - **curling** American Fitness - Find Articles - [Traduzir esta página]
The roarin game - **curling** from American Fitness in Array provided free by **...** clubs were formed in the **Scandinavian countries**, the Alps, New Zealand, **...**
findarticles.com/p/articles/mi_m0675/is_n2_v8/ai_8853209 - 29k -
Em cache - Páginas semelhantes

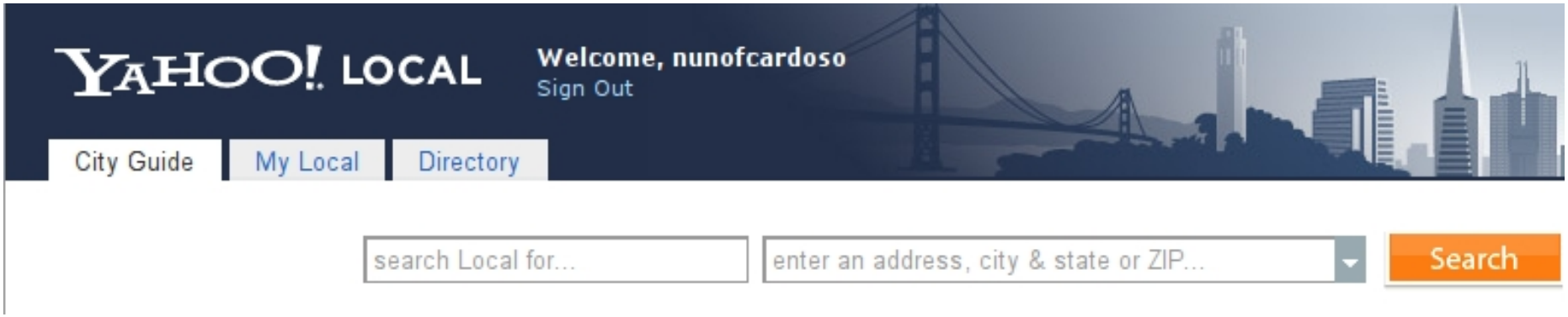
Links Patrocinados

Scandinavian clubs
Connections worldwide. Join free
Totally anonymous. Meet now.
ScandinaviaSingles.com

Google search for “Curling **IN** Scandinavian countries”

Web search engines go local...

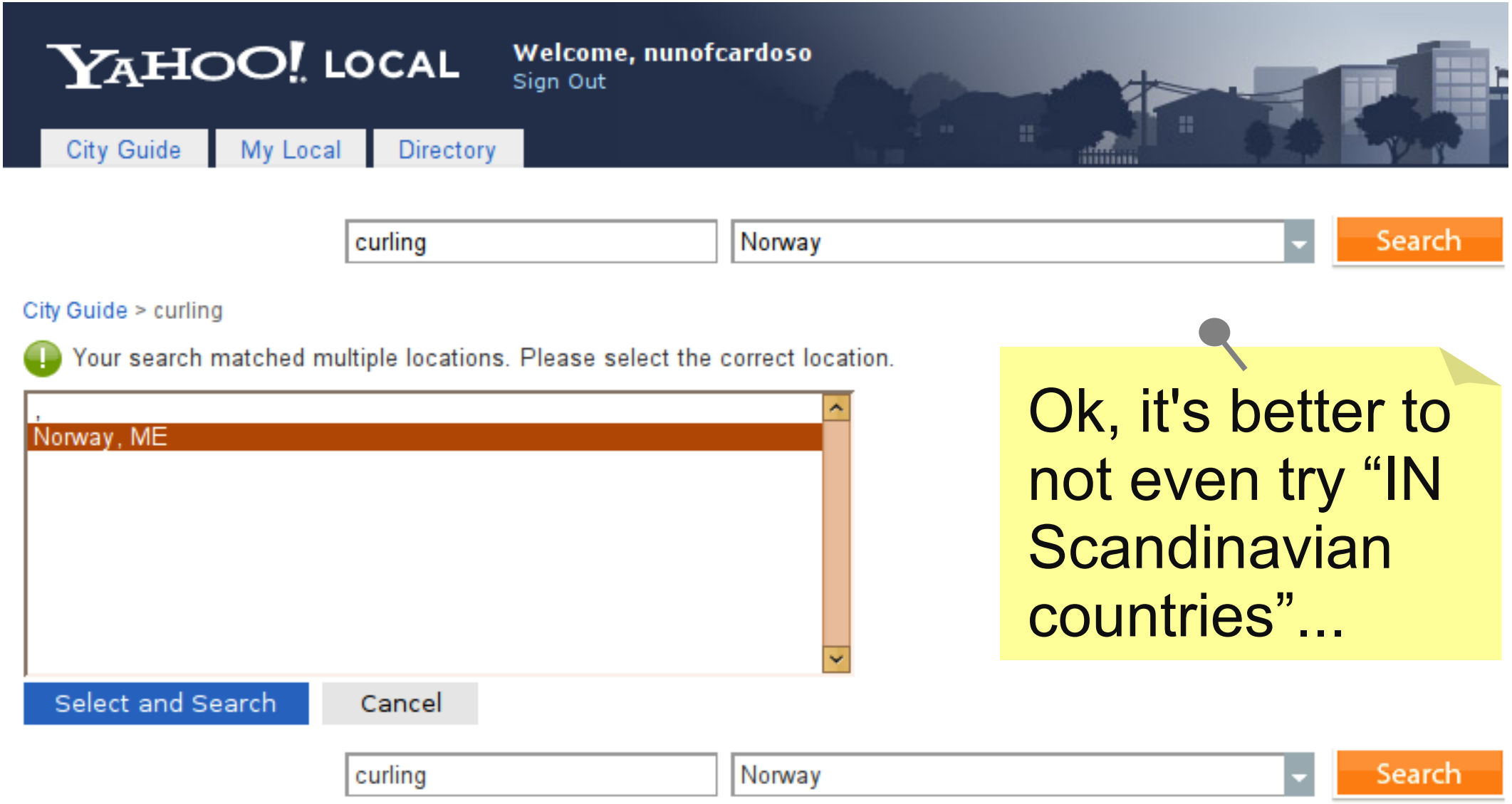
- Local Yahoo! (local.yahoo.com)



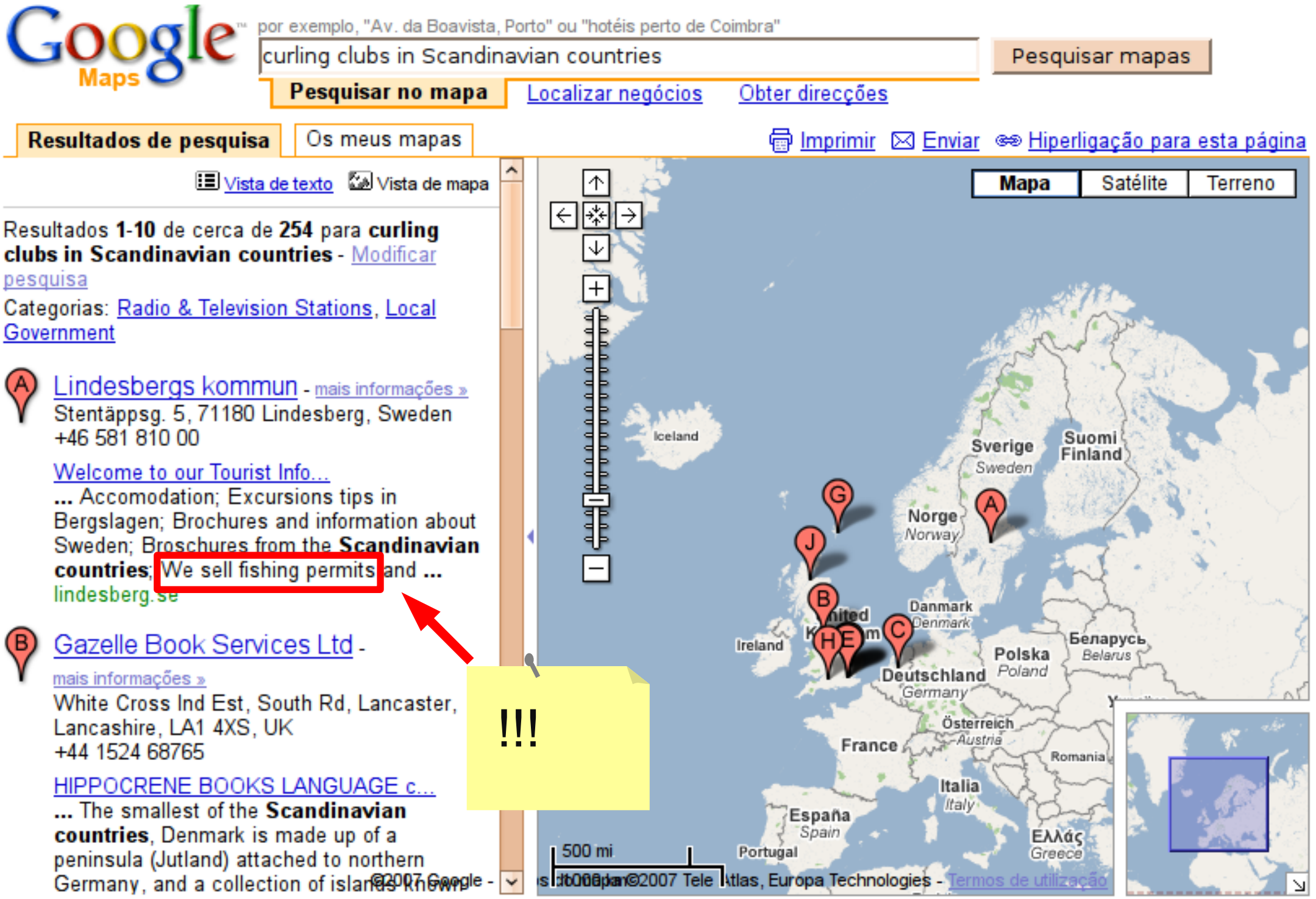
- Local Google (= Google Maps)

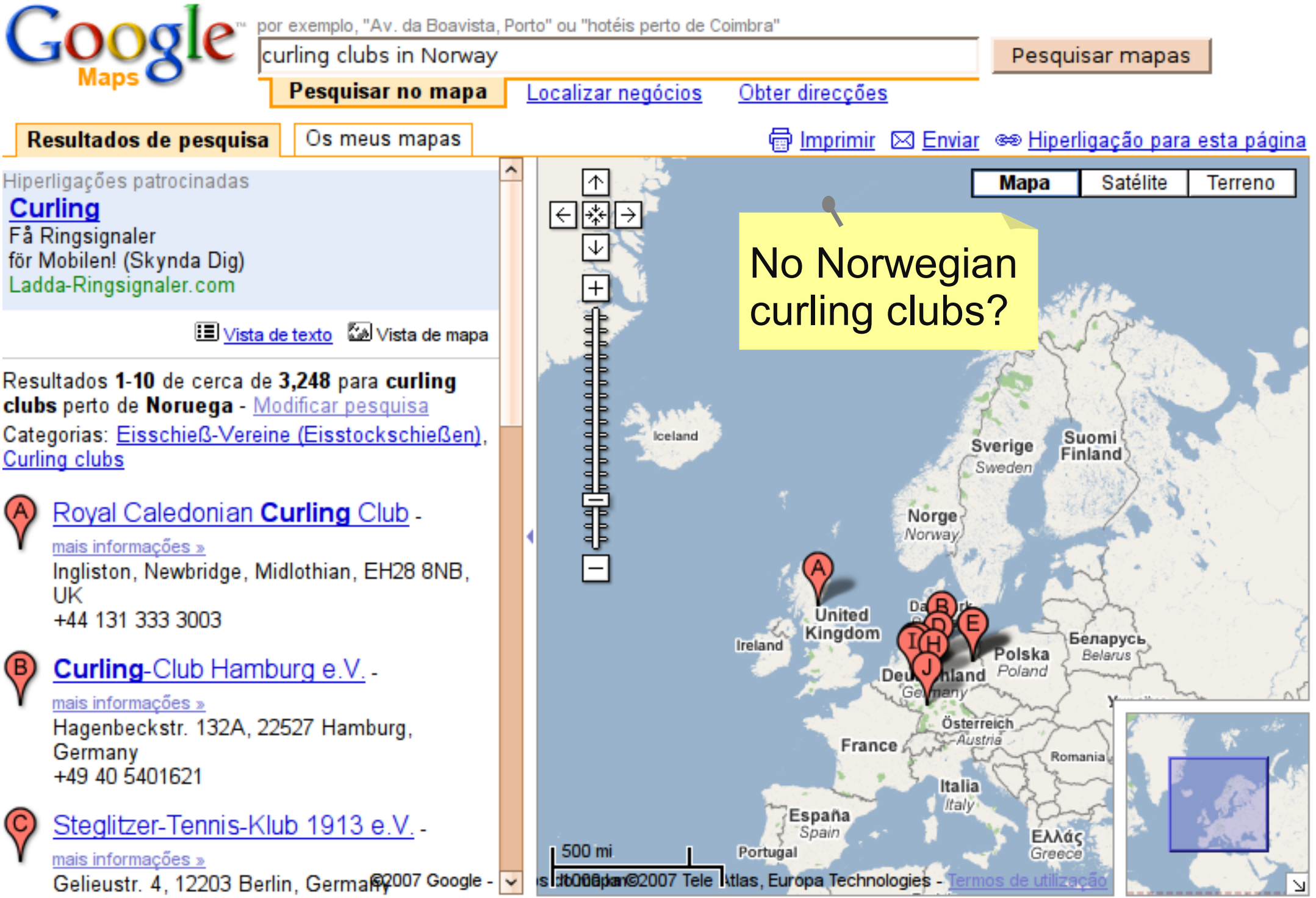
How well are they handling our geographic restrictions?

Curling in Norway, Yahoo! local...



Ok, it's better to not even try “IN Scandinavian countries”...


Google Maps
por exemplo, "Av. da Boavista, Porto" ou "hotéis perto de Coimbra"
curling clubs in Scandinavian countries
Pesquisar mapas
Pesquisar no mapa
Localizar negócios
Obter direcções
Resultados de pesquisa
Os meus mapas
Imprimir
Enviar
Hiperligação para esta página
Vista de texto
Vista de mapa
Resultados 1-10 de cerca de 254 para curling clubs in Scandinavian countries - Modificar pesquisa
Categorias: Radio & Television Stations, Local Government
A
Lindesbergs kommun - mais informações »
Stentäppsg. 5, 71180 Lindesberg, Sweden
+46 581 810 00
Welcome to our Tourist Info...
... Accomodation; Excursions tips in Bergslagen; Brochures and information about Sweden; Broschures from the Scandinavian countries; We sell fishing permits and ...
lindesberg.se
B
Gazelle Book Services Ltd -
mais informações »
White Cross Ind Est, South Rd, Lancaster, Lancashire, LA1 4XS, UK
+44 1524 68765
HIPPOCRENE BOOKS LANGUAGE c...
... The smallest of the Scandinavian countries, Denmark is made up of a peninsula (Jutland) attached to northern Germany, and a collection of islands known
!!!
Mapa
Satélite
Terreno
Iceland
Sverige
Sweden
Suomi
Finland
Norge
Norway
Danmark
Denmark
Ireland
Беларусь
Belarus
Polska
Poland
Deutschland
Germany
Österreich
Austria
France
Romania
Italia
Italy
España
Spain
Portugal
Ελλάς
Greece
500 mi
©2007 Google - Dados do mapa ©2007 Tele Atlas, Europa Technologies - Termos de utilização


Google Maps
por exemplo, "Av. da Boavista, Porto" ou "hotéis perto de Coimbra"
curling clubs in Norway
Pesquisar mapas
Pesquisar no mapa
Localizar negócios
Obter direcções
Resultados de pesquisa
Os meus mapas
Imprimir
Enviar
Hiperligação para esta página
Hiperligações patrocinadas
Curling
Få Ringsignaler
för Mobilen! (Skynda Dig)
Ladda-Ringsignaler.com
Vista de texto
Vista de mapa
Resultados 1-10 de cerca de 3,248 para curling clubs perto de Noruega - Modificar pesquisa
Categorias: Eisschieß-Vereine (Eisstockschießen), Curling clubs
A Royal Caledonian Curling Club -
mais informações »
Ingliston, Newbridge, Midlothian, EH28 8NB, UK
+44 131 333 3003
B Curling-Club Hamburg e.V. -
mais informações »
Hagenbeckstr. 132A, 22527 Hamburg, Germany
+49 40 5401621
C Steglitzer-Tennis-Klub 1913 e.V. -
mais informações »
Gelieustr. 4, 12203 Berlin, Germany
©2007 Google -
Mapa
Satélite
Terreno
No Norwegian curling clubs?
Iceland
Sverige
Sweden
Suomi
Finland
Norge
Norway
United Kingdom
Ireland
Polska
Poland
Беларусь
Belarus
Österreich
Austria
France
Romania
Italia
Italy
España
Spain
Portugal
Ελλάς
Greece
500 mi
©2007 Tele Atlas, Europa Technologies -
Termos de utilização

Motivation for GIR

- Users have information needs that have a given **scope of interest**.
- Present search engines do not understand **placenames** in queries as a different kind of relevance criteria.
- Documents should be retrieved according to **topic relevance** and **geographic relevance**.
- IR systems must adapt to the users, and not the users to the IR system!

2) GIR approaches

- Overview of IR and GIR
- The case of GeoTumba

How does a typical web **IR** system work?

1) Crawling: downloads “all” documents from the *web*;

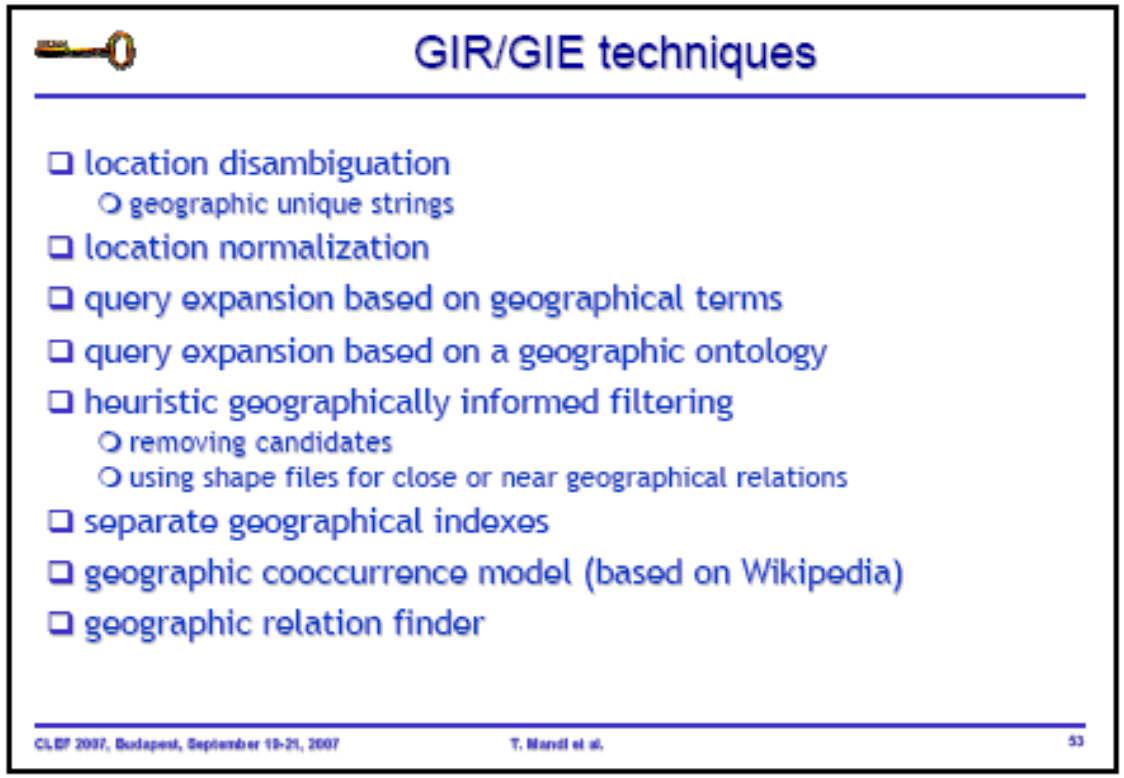
2) Storage: pre-processes and stores documents locally;

3) Indexing and **Ranking:** generates term indexes and weighs documents;

4) Interface: processes queries and presents results to the user.

How does a web **GIR** system work?

- ... can't tell! GIR is a recent area:
  - First workshop: Analysis of Geographical References, hosted in NAACL-HLT in 2003.
  - Annual GIR workshops started in 2004.
  - Evaluation contests started in 2005 (GeoCLEF).
- Many systems with different approaches.
  - GeoCLEF 2007: 13 participants, 8 approaches!



Mandl et al, ”Overview of GeoCLEF 2007”

The case of Geotumba

Pesquisa Geográfica na Web Portuguesa

- **GeoTumba** is a web search engine with geographic capabilities developed by the University of Lisbon.
- Research Team: Mário J. Silva (head), Ana Paula Afonso, Bruno Martins, David Cruz, Marcirio Chaves, Nuno Cardoso and Paulo Pombinho, Leonardo Andrade and Sérgio Freitas
- GREASE project since January 2004.
  - Diana Santos and Chris Jones (advisors)
  - Collaboration with SINTEF/Linguateca.

Geotumba

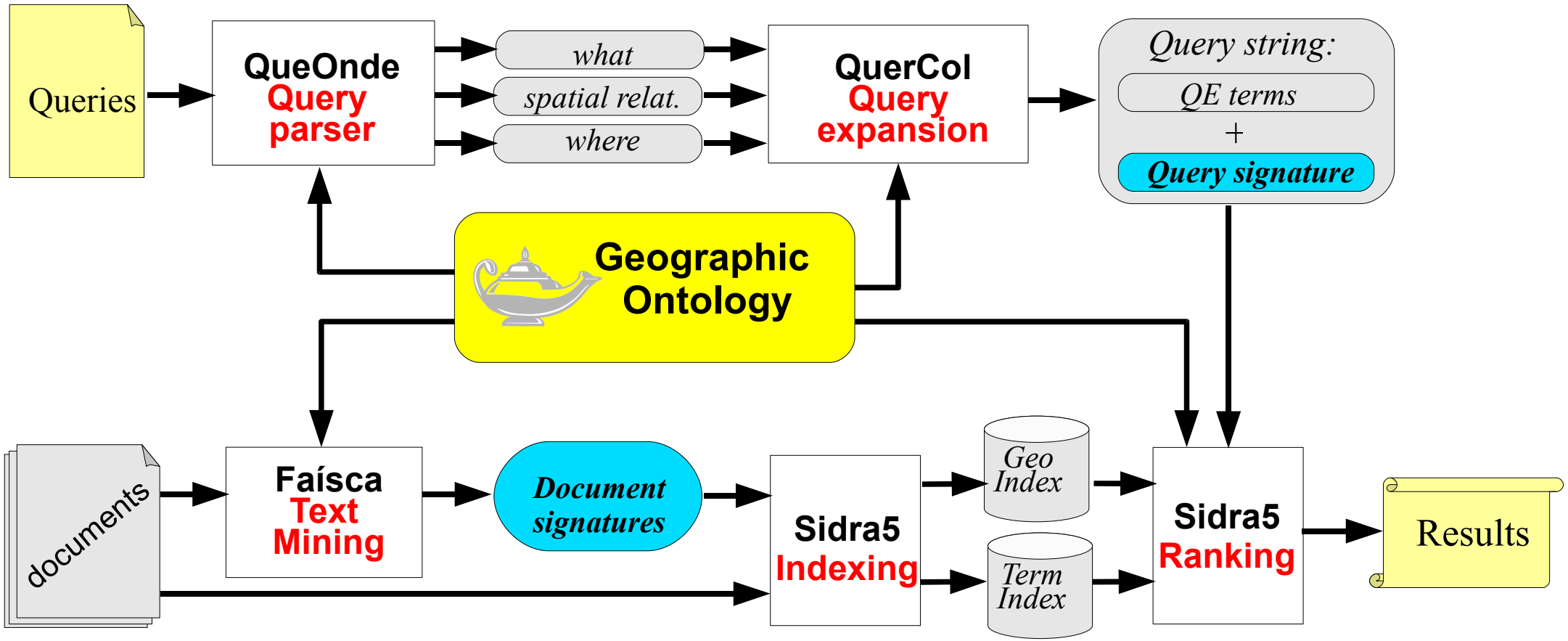
Pesquisa Geográfica na Web Portuguesa

Geotumba's main characteristics:

- **Geographic knowledge:** provided by a geographic ontology, GeoNET PT 01.
- **Text mining** for placenames, **match** them into **features** of the ontology.

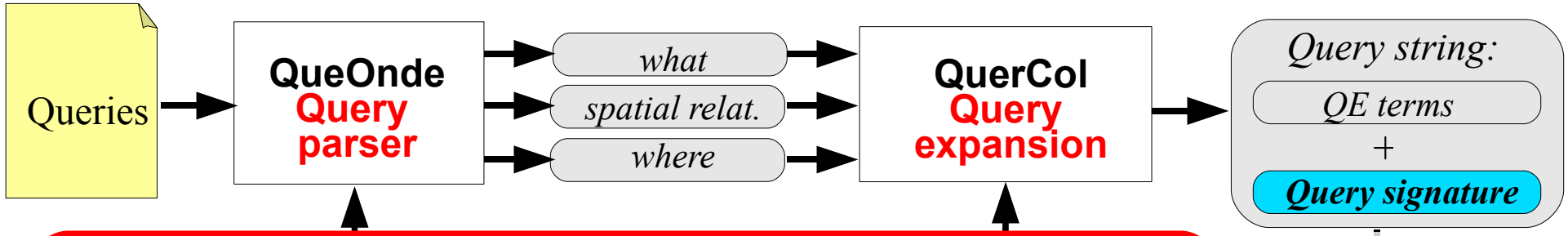
- Assign a **list of features** to each document, the document geographic signature.
- The non-geographic part (*curling*) is ranked by text weighting scheme BM25; geographic part (*Norway*) is ranked by a set of heuristics.
- Doc. ranking score: 0.5 non-geo + 0.5 geo.

GIR system (2007)



GIR system (2007)

1. Query parsing



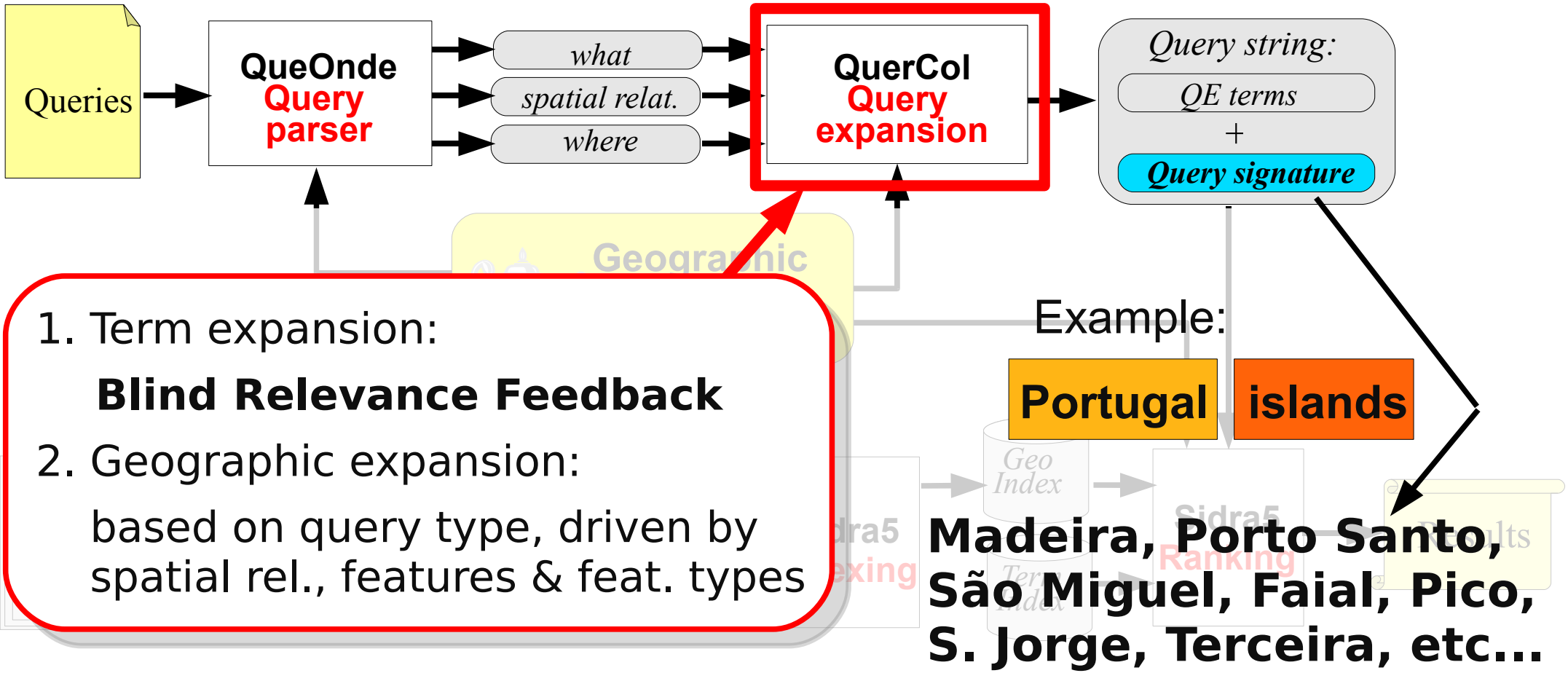
- **Splits into <what, spat.rel., where>**
- **Recognizes *features* & *feature types***

Example: Sea traffic in Portuguese islands =

Sea traffic | in | Portugal | islands

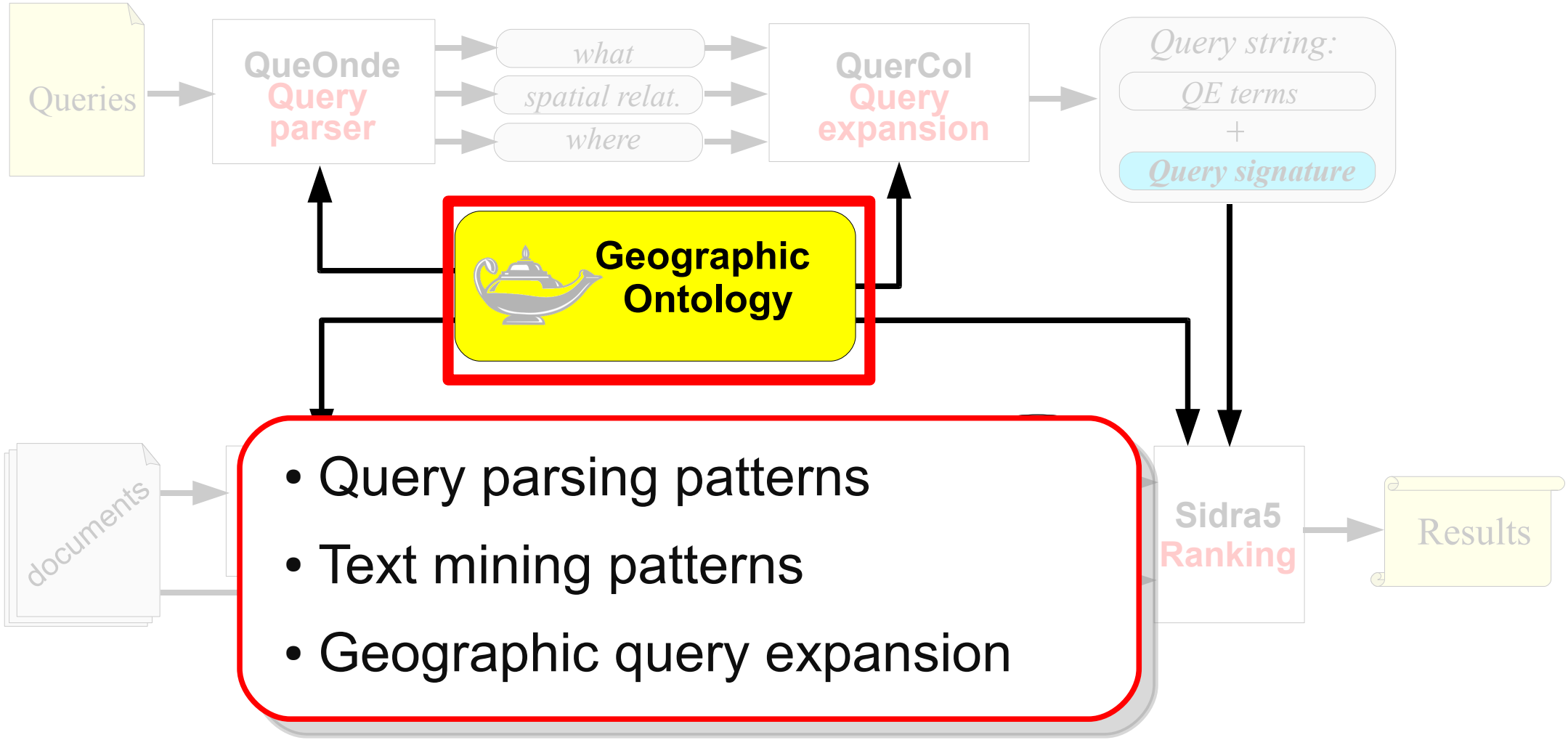
GIR system (2007)

2. Query expansion



GIR system (2007)

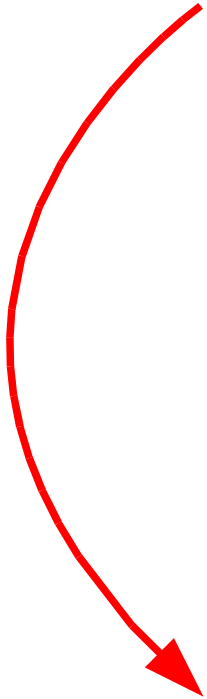
3. Geographic knowledge



- Query parsing patterns
- Text mining patterns
- Geographic query expansion

Our geographic ontology

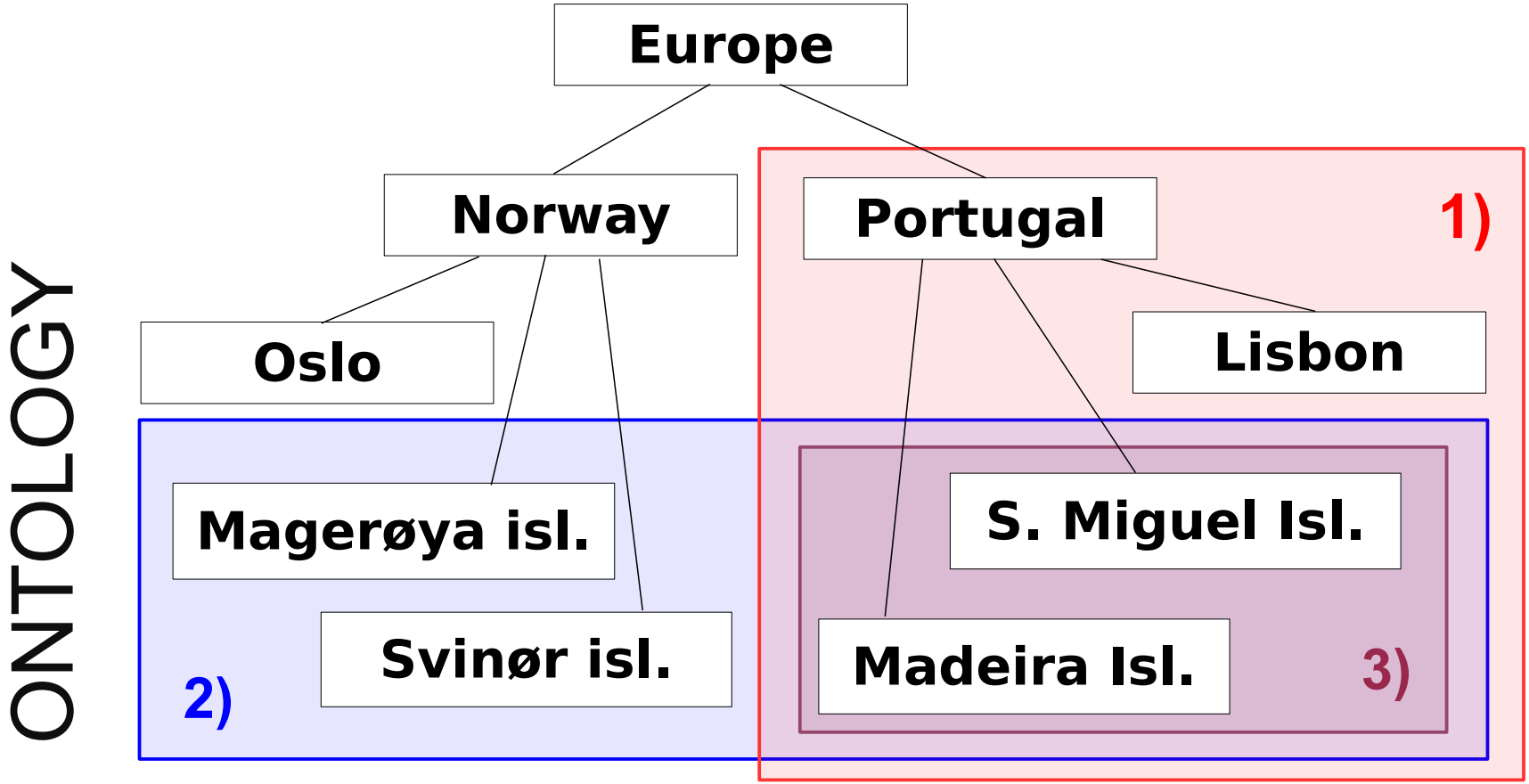
- Encodes geographic knowledge and provides it to GIR components.
- Contains **features**, **feature types** and **relationships** among them.



Physical Domain				Administrative Domain	
Island	205	Sea	5	Place	4023
Airport	107	Cathedral	3	ISO-3166-2	3976
River	86	Ocean	2	Administrative division	3212
Mountain	85	Mountain Range	2	Agglomeration	751
Lake	66	Strait	1	ISO-3166-1	239
Circuit	63	Channel	1	Capital city	233
Region	23	Planet	1	**Total**	12434
Continent	7	**Total**	657		
Names			14408	Centroids	4204
Features			13091	Bounding boxes	2083
Feature Types			21	*adjacent* relationships	11307
				part-of relationships	13762

Mapping placenames into features and feature types

1) *sea traffic* **in Portugal**
2) *sea traffic* **in islands**
3) *sea traffic* **in Portuguese islands**



GIR open questions...

- How to **represent** geographic knowledge?
- How to **use** it properly on GIR components?
- How to handle **ambiguity** of placenames in documents and queries?
- How to **classify** documents geographically?
- How to measure the **geographic affinity** between queries and documents?
- How to **present** the results to the user?

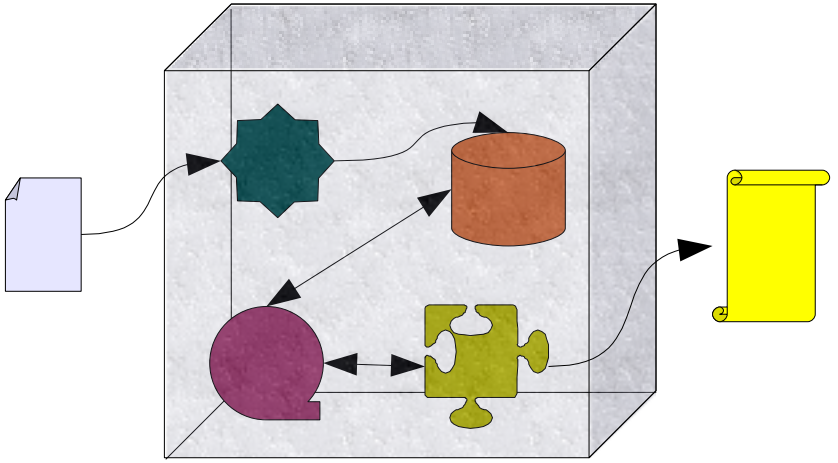
3) GIR Evaluation

- In-house evaluation of GIR components
- International evaluation contests: GeoCLEF

Evaluation of GIR systems

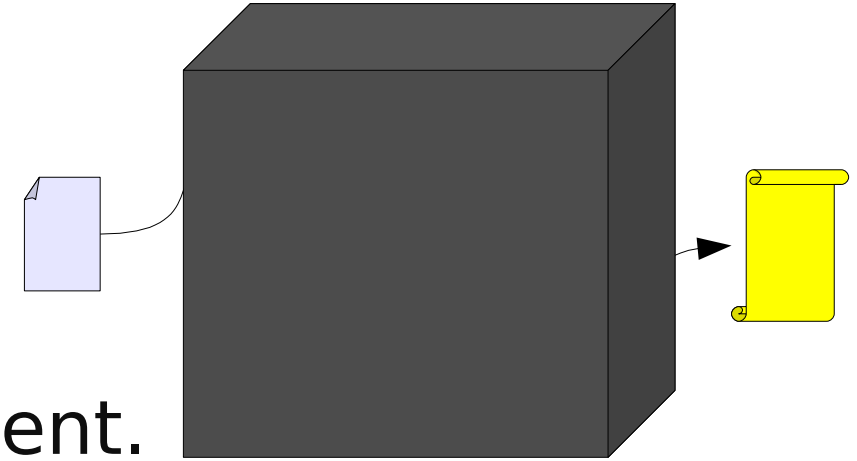
1) In-house evaluation

- test each GIR component in a “glass-box” style evaluation.
- analyse points of failure, measure performance of each component.



2) Evaluation contests

- “black-box” style evaluation.
- compare different systems, several approaches.
- common evaluation environment.



GeoCLEF evaluation setup

- Text collections: newspapers in 3 languages
  - Portuguese collection provided by Linguateca.
- Every year, 25 topics released in several languages.
- Relevance judgement made by humans
- Measure: Prec., Rec., MAP, ...

<top>
<num>10.2452/51-GC**</num>**
<title>Oil and gas extraction found between the UK and the European Continent**</title>**
<desc>To be relevant documents describing oil or gas production between the UK and the European continent will be relevant**</desc>**
<narr>Oil and gas fields in the North Sea will be relevant.**</narr>**
</top>

Our evaluation results so far

- XLDB participated in GeoCLEF 2005, 2006 and 2007 (since its beginning).
- Main challenge: outperform IR.
- So far, GIR failed to outperform IR (same results for other groups).
- What are we doing wrong?

  approach? query processing? text mining? geographic ranking? scope assignment?

➡ Only detailed evaluation will tell...

XLDB in GeoCLEF 2007

- IR outperformed by a IR/GIR mixed strategy.
- Our GIR only approach still worse than IR!
- In progress:
  - Detailed analysis for each of the 25 topics.
  - Thorough evaluation on each GIR component: query parsing, query expansion, text mining, index & ranking.

4) My PhD Research Plan

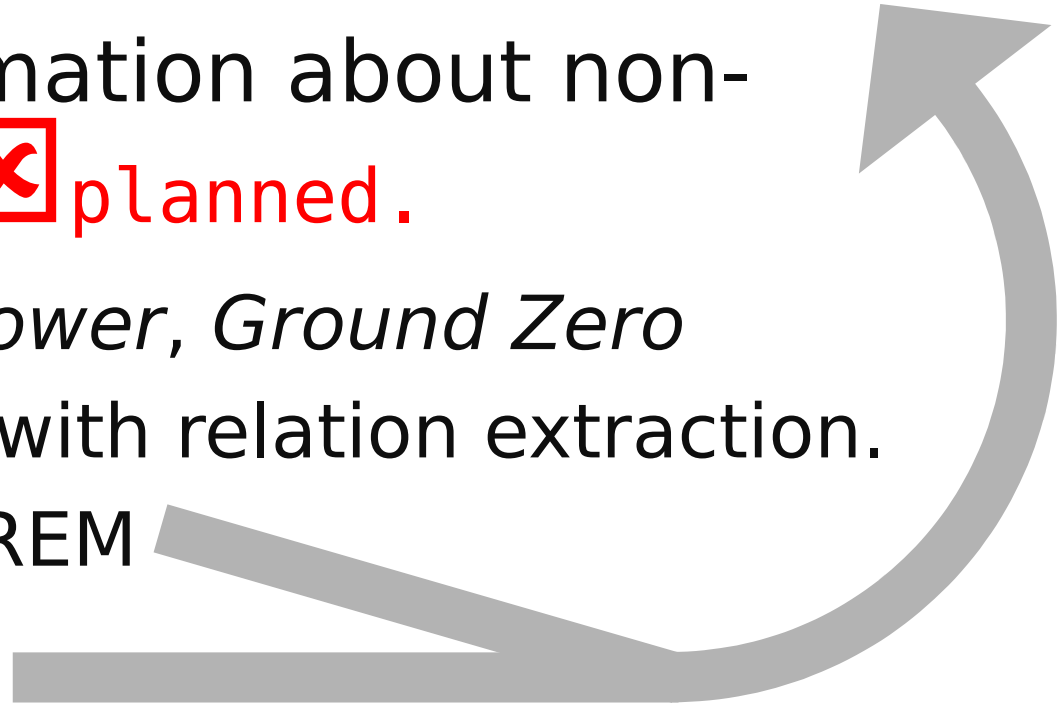
PhD Grant

“Automatic Query Reformulation for Web Geographic Search Engines”

- PhD advisors:
  - Diana Santos (Linguateca/SINTEF ICT)
  - Mário J. Silva (University of Lisbon)
- Started 1st March 2007
- 3 year scholarship

Milestones 2007/2008

- 1st step: GeoCLEF experiments. ☑ done.
- 2nd step: Evaluation. ◉ ongoing.
  - a) Detailed error analysis of each topic.
  - b) development of a good evaluation environment.
- 3rd step: Acquiring information about non-standard placenames. ☒ planned.
  - Example: *SINTEF*, *Eiffel Tower*, *Ground Zero*
  - implies new NER system with relation extraction.
  - Evaluate it in Second HAREM
- 4th Step: GeoCLEF 2008



2a) Kinds of possible errors

Off-line processing:

- **Document processing:** failure to capture geographic evidence (placenames and relations), failure to reason into ontology features.
- **Document Geographic Signatures:** erroneous ranking of irrelevant geographic evidence.

2a) Kinds of possible errors

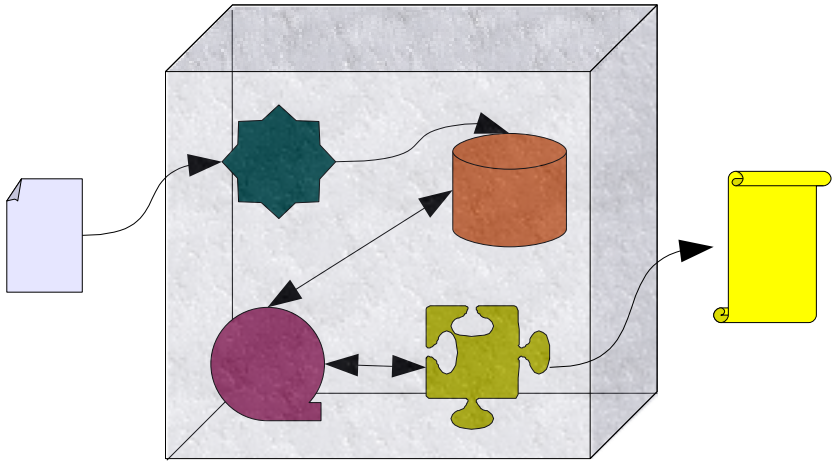
On-line:

- **Query parsing:** failure to detect relevant terms, spatial relationship, feature and feature types.
- **Term Expansion:** adding unrelated terms, skipping related terms.
- **Ontology Processing:** failure to translate query scope into a list of features
- **Document Ranking:** failure to bring relevant documents to the top of the result list.

Analyse one topic and identify points of failure.

2b) In-house evaluation environment

- **Input:** topics/queries and human relevance judgements.
- **Output:** evaluation reports.
- **Desiderata:**
  - Measure each component separately.
  - Logging and debugging capabilities for each point of failure.
  - Invoke components under controlled scenarios.



2b) In-house evaluation environment

Evaluation for different scenarios:

- **Topic by topic** – debug a particular component or a query.
- **By topic type** – measure system's ability to deal with a particular type of input.
  - See Santos & Chaves (2005) type classification
- **Overall** – measure global performance and contribution of each component to the overall results.

3) non-standard placenames

- Ex: *SINTEF*, *Eiffel Tower*, *Ground Zero*
- Develop a new NER system, **REMBRANDT**
  - Detection of placenames in context
    - Why? Try “Norwegian Wood”
  - Exploit Wikipedia's link structure, category and annotation metadata, and entity relation detection.
- Participation in Second HAREM (March 2008)
  - Independent evaluation contest, with already manually annotated corpora

4) GeoCLEF 2008

- Improve GIR system with REMBRANDT
- Use evaluation environment to measure impact on performance
- Participation in GeoCLEF 2008 (May/June 2008)
  - Special sub-task for Wikipedia

For GeoCLEF 2008:

- Pilot subtask: use Wikipedia snapshot as document collection:
  - Richer resource for documents with geographic content;
  - allows different types of topics.
- Documents about “Riots in Los Angeles” is more likely to be found in newspapers...
- ... but documents about “Vineyard areas near European rivers” more likely to be reported in Wikipedia.

Epilogue

- GIR systems are useful and challenging
- Still on its early steps, still not clear what are the most promising approaches
- More message understanding & geographic reasoning must be made carefully
- How to measure progress: evaluation, evaluation, evaluation.

Geographic IR Challenges

The End.

by Nuno Cardoso

Faculty of Sciences,
University of Lisbon, LASIGE

Presentation held at SINTEF ICT, Oslo, Norway, 4th December, 2007

Entity Relation Extraction for placenames

- **Inclusion**: (ex: “Adelaide, Australia”): Australia suggests that Adelaide is a placename, not a proper name
  - *partOf(Adelaide, Australia)*
- **Grounding**: (ex: “SINTEF ICT is in the outskirts of Oslo”) - now, I know that documents mentioning SINTEF ICT may have Oslo as a potential scope of interest (Ex: “research units near Oslo”).
  - locatedIn(*Sintef ICT, Oslo*)
- **Alternative**: (ex: “California as CA”)
  - Alternative(*CA, California*)