

A GIR Architecture with Semantic-flavored Query Reformulation

Nuno Cardoso and Mário J. Silva

Overview

1. Motivation

2. Reasoning towards the proposed GIR architecture

3. GIR Architecture Overview

a) Handling queries

b) Handling documents

c) Retrieving documents

4. Prototype development status and snapshots

5. Current challenges



Motivation

- Queries have *entities*, and entities have *semantic information*.
- Term-statistic query reformulation works at *term level*, not entity level.



Term

Entity



Motivation (cont.)

- In order to faithfully reformulate queries regarding the user information need, we need a semantic query reformulation approach that uses such **semantic information from entities and their relationships** in its reformulation strategy.
- To validate this theory, let's apply it to a **GIR prototype** and measure its performance over geographic queries

Reasoning towards the proposed GIR architecture

- 1. To perform semantic query reformulation, we need to **detect** and **ground entities**, and have access to more **information** about them;
- 2. To do that, we need a **knowledge base of entities** and easy access to **third party knowledge bases** (Wikipedia, DBpedia, geographic ontologies, etc);



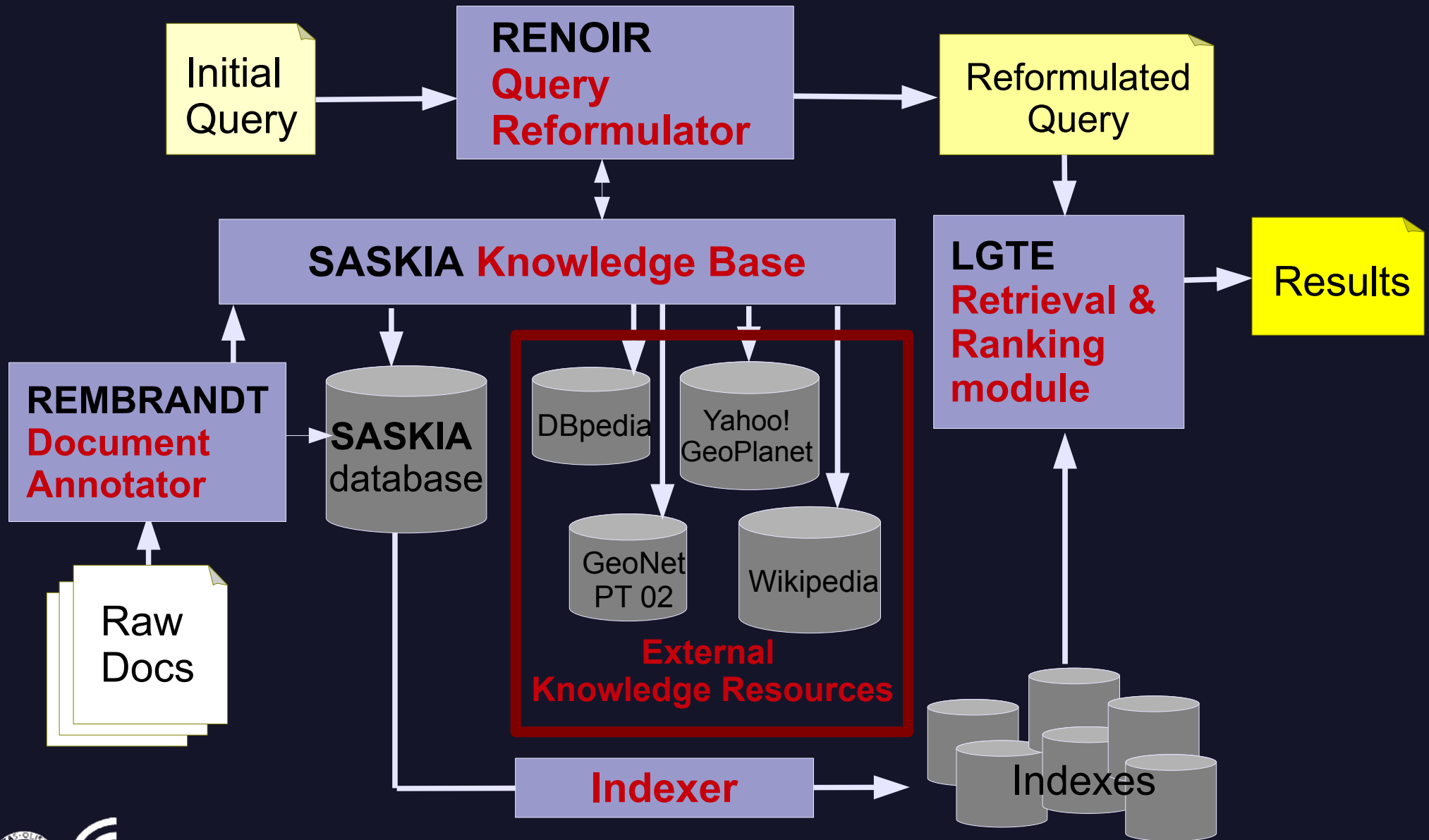
Requirements for the proposed GIR architecture (cont.)

- 3. Once query **entities** are grounded and the **information need** is captured, the retrieval phase should adapt to each query types, and use this information to rank documents (or “*one size doesn't fit all*”);
- 4. We need a retrieval & ranking module that weights documents regarding its **terms**, **entities**, **geoscope** and **temporal scope**;

Requirements for the proposed GIR architecture (cont.)

- 5. Finally, we need **all collection documents** to have their entities, geoscope and temporal scope **grounded** and **indexed**.

GIR Architecture



Handling queries

1. RENOIR grounds all entities and detect the user intentions – geoscope, expected answer type (EAT), etc

“ **Music festivals** in **Germany** ”

Role: **EAT**

Ground: **DBpedia:Category:Music_festivals**

Role: **Geoscope**

Ground: **DBpedia:Germany_(Country)**
WOEID: 23424829

About: http://dbpedia.org/resource/Category:Music_festivals
An Entity in Data Space: dbpedia.org

Property	Value
rdf:type	▪ skos:Concept
rdfs:label	▪ Music festivals
skos:broader	▪ dbpedia:Category:Live_music ▪ dbpedia:Category:Music_events ▪ dbpedia:Category:Festivals
skos:prefLabel	▪ Music festivals
is skos:broader of	▪ dbpedia:Category:International_music_festivals

```
--<place yahoo:uri="http://where.yahooapis.com/v1/place/23424829" xml:lang="en">  
<woeid>23424829</woeid>  
<placeTypeName code="12">Country</placeTypeName>  
<name>Germany</name>  
<country type="Country" code="DE">Germany</country>  
<admin1/>  
<admin2/>  
<admin3/>  
<locality1/>  
<locality2/>  
<postal/>  
--<centroid>
```

Handling queries (cont.)

2. Use the SASKIA knowledge base to add **concrete answers**, use them as expanded terms for selected indexes

Initial
query

term: music festivals in Germany



Final
query

term: music festivals germany “wacken open air”
“zappanale” “summerjan” “summer breeze open air”
event: “Wacken Open Air” “Zappanale” “Summerjan”
“Summer Breeze Open Air”
place: Germany
geoscope: Germany@WOEID-23424829

Document
tagged by
REMBRANDT

Handling documents

Wacken Open Air takes place annually in the small town of **Wacken**, in **Germany**. **Wacken** is a metal festival.

1. REMBRANDT recognizes all **named entities** from documents, grounds them to **entities**, maps places to **geoscopes**, stores everything into SASKIA database.



Handling documents (cont.)

2. Index generation. There are 3 kinds of indexes:

- a) **Term index** – classic inverted index for document terms;
- b) **Named entity index** - inverted index for named entity terms, one for each HAREM's* semantic classification;
- c) **Signature index** - indexes of geographic and temporal signatures of documents.



*HAREM is a evaluation contest for named entity recognition systems. More information in <http://www.linguateca.pt/HAREM/>

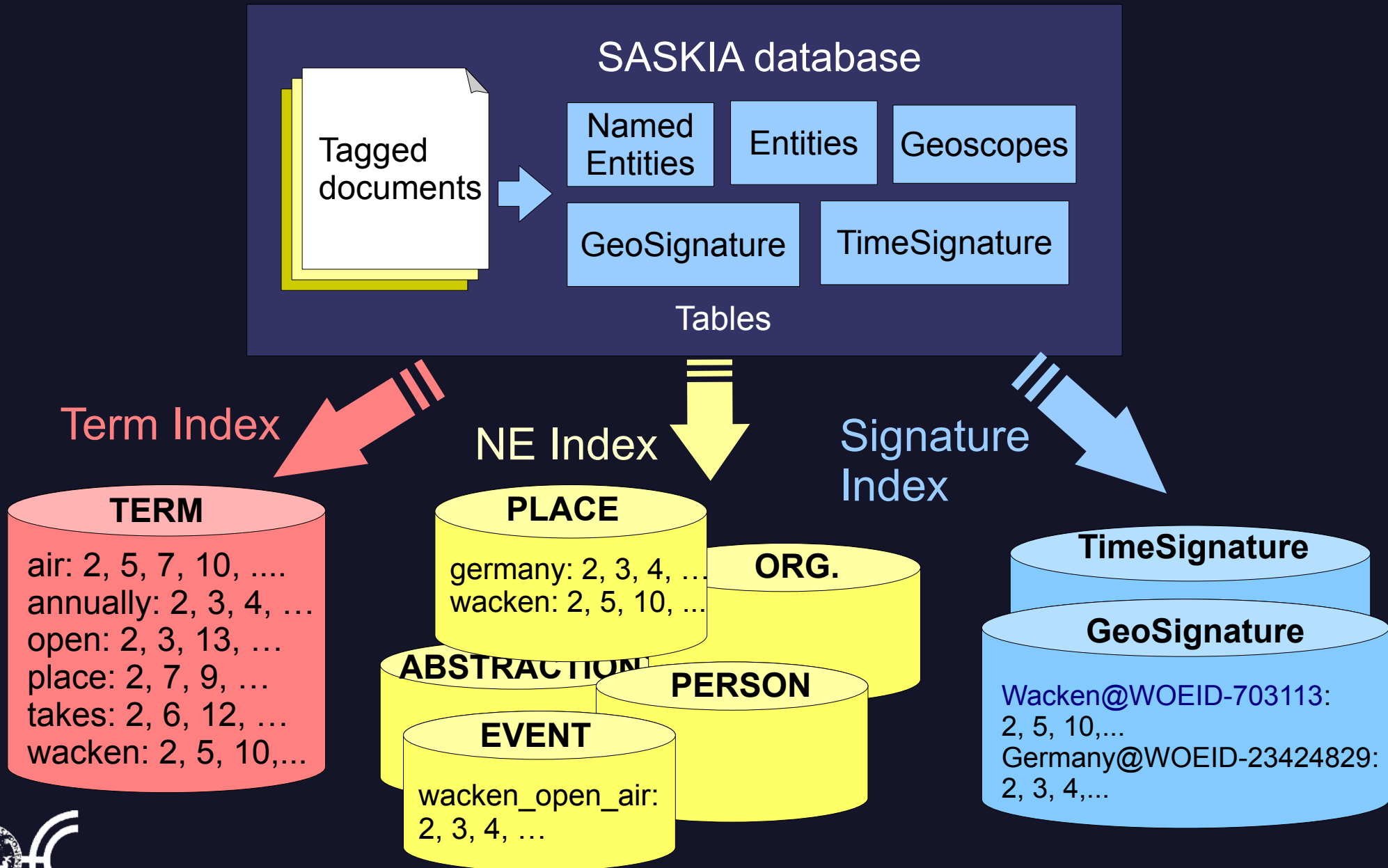
Document signatures

Surrogate of entities referred in the document that describe its scope.

```
<GeoSignature totalcount="4">
  <Doc id="571048" lang="en">
    <Place count="2" woeid="1467052">
      <NE>Harare</NE>
      <Entity>Harare</Entity>
      <Type>@HUMAN</Type>
      <Subtype>@DIVISION</Subtype>
      <DBpediaClass>Area</DBpediaClass>
      <Ancestor>Harare</Ancestor>
      <Ancestor>Harare</Ancestor>
      <Ancestor>Zimbabwe</Ancestor>
    </Place>
    <Place count="2" woeid="23425004">
      (...)
    </Place>
  </GeoSignature>
```

```
<TimeSignature>
  <Doc id="523634" lang="en">
    <Time count="1">
      <NE id="3645">2006</NE>
      <TG>!:Y+2006</TG>
      <Index>2006</Index>
    </Time>
    <Time count="1">
      <NE id="3646">27th May, 2006</NE>
      <TG>!:Y+2006M05S27</TG>
      <Index>20060527</Index>
    </Time>
    (...)
  </TimeSignature>
```

Handling documents (cont.)



Retrieving documents

RENOIR's reformulated query

term: music festivals germany
"wacken open air" "zappanale"
"summerjan" "summer breeze
open air"

event: "Wacken Open Air"
"Zappanale" "Summerjan"
"Summer Breeze Open Air",
place: Germany

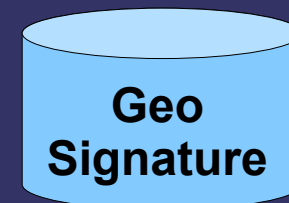
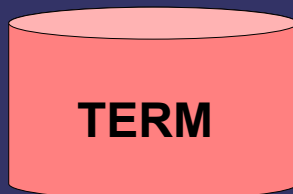
geoscope:
Germany@
WOEID-23424829

Lucene with
GeoTemporal
Extensions



Results


REMBRANDT / SASKIA
indexes




Prototype development status

- Prototype online at <http://xldb.di.fc.ul.pt/Rembrandt/>
- Four collections (PT and EN) annotated, indexed and available to be queried.
- Lucene LGTE looks fast and stable
- Some user interface add-ons (NE tags, document details, etc)
- RENOIR is still under development, tough :(

Prototype snapshots

Português | English


[Home](#) > [REMBRANDT Web Service](#) > [RENOIR QA service](#)

 Suggestions User: [Nuno Cardoso](#)
Collection: [New York Times 2002-2005](#)

Results 1 - 10 of 8935 for **forest fires in Portugal**. [More info >>](#)

Results

- 1** [NEW FOREST MANAGEMENT RULES DRAW PRAISE AND CRITICISM](#)
the nation 's 155 national forests and making it easier for regional forest managers to decide whether(...) forest and grasslands . They also cut back on requirements for public participation in forest planning decisions . Forest Service officials said the rules are designed to give local foresters more(...) foresters see as a legal and regulatory gridlock that has delayed forest plans for years because(...) want the forest to look and be in the future , " said Rick D. Cables , the Forest Service
3KB - 2004-12-22 | [Show](#) | [Details](#) | [More info](#)
- 2** [FOREST DISSERVICE \(FOR USE FOREST DISSERVICEUNTIL NOW](#) , both national forests in New England have held the line against the all-terrain vehicles that chew up the forest floor , speed erosion , spread invasive species(...) Forest in New Hampshire and Maine maintains the ban on ATVs . But the recently released draft plan for the Green Mountain National Forest in Vermont foresees letting ATV owners cross national forest land on " corridors " connecting with a larger trail system outside the national forest .
National
5KB - 2005-04-24 | [Show](#) | [Details](#) | [More info](#)
- 3** [AFTER THE STORMS CLEARING OF FOREST READY TO RESUME](#)
new plant growth and renewal . A lack of fires in the forest created a dense canopy that allowed relatively little sunlight to reach the forest floor , discouraging many plants from growing(...) habitat at Lake Wales Ridge State Forest . Now , those plans are back on track . A salvage logging operation to remove large trees from the floor of the 26,488-acre state forest in eastern Polk County(...) so forest visitors will know what to expect . Signs will alert visitors to what 's happening . " We




A map showing search results locations. The map includes North America, Europe, and South America. There are several red location pins with numbers: 8 in North America, 2 in Europe, 6 in Europe, 9 in North America, and 8 in North America. The map is powered by Google and includes a search bar and navigation controls.

Prototype snapshots

REMBRANDT Português | English

Home > REMBRANDT Web Service > RENOIR QA service

 Suggestions

[Advanced search >>](#)

Results 1 - 10 of 8935 for Forest Fires in Portugal. [More info >>](#)

Results **SPAIN SCOLDS CARELE**

Statistics for document

Document details
Creation date: 2005/Set/17
Tagged date: 2009/Dez/06
Tagged with REMBRANDT v.43:1.0-b1427


Sentences and NEs
Number of sentences in title: 1
Number of sentences in body: 29
Number of NEs in title (document): 0
Number of NEs in body (document): 5
Number of NEs in title (pool): 1
Number of NEs in body (pool): 34

Top 10 NEs
7 - **Spain** @LOCAL @HUMANO @PAIS
2 - **Portugal** @LOCAL @HUMANO @PAIS
1 - **July , 11** @TEMPO @TEMPO_CALEND @DATA
1 - **Mediterranean** @LOCAL @FISICO @AGUAMASSA

Geographic signature
<GeoSignature version="1.0" totalcount="3">
<Doc id="834259" original_id="NYT_ENG_20050917" lang="en" />
<Place count="2" woeid="23424925">
<NE id="1016436">Portugal</NE>
<Name>Portugal</Name>
<Type>@HUMANO</Type>
<Subtype>@PAIS</Subtype>

Time signature
<TimeSignature version="1.0" totalcount="1">
<Doc id="834259" original_id="NYT_ENG_20050917" lang="en" />
<DocDateCreated>20050917</DocDateCreated>
</TimeSignature>

Map



POWERED BY Google

Dados do mapa ©2010 Tele Atlas, Europa Technologies - [Termos de utilização](#)

Generated in 2010/Fev/14

Current challenges

RENOIR:

- **Rule set chains** for entity detection in queries for simple and complex queries
- **Picking the best strategy** for query reasoning:
 - What knowledge resources should we use?
 - Mapping relationships to DBpedia properties (ex: “born in” → dbpedia-owl:birthplace)
 - Handle low recall results (if “*Romanian writers born in Bucharest*” returns few or no answers, is there a plan B?)



Current challenges (cont.)

REMBRANDT / SASKIA:

- Tagging documents and populating DB is slow, complex and requires supervising
- Example: NYT collection (2002-2005)

Nr of documents	315.371
Nr of named entities	17.952.142
Nr of classifications assigned for named entities	18.364.572
Nr of classifications grounded to entities	3.344.235
Nr of classifications grounded do a place	588.621
Nr of docs with non-empty GeoSignature	202.624 (64%)
Nr of docs with non-empty TimeSignature	70.403 (22%)
Total entities in Saskia DB	37.001
Total geoscopes in Saskia DB	8.741

In conclusion...

- **Semantic query reformulation** requires many changes in a GIR system
- A **prototype is now online** (currently with limited functionalities on query handling...)
- The **prototype is available** (web-service, API, or download) for other researchers (GIR et al.)
- **Tasks**: have RENOIR ready, evaluate the query reformulation effect on results, use GeoCLEF benchmark to compare with other QR approaches

The end

A GIR Architecture with Semantic-flavored Query Reformulation

Nuno Cardoso and Mário J. Silva

Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE

ncardoso@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt