



FACULDADE DE CIÊNCIAS | UNIVERSIDADE DE LISBOA

# Experiments with Semantic- favored Query Reformulation of Geo-Temporal Queries

Nuno Cardoso<sup>1</sup> and Mário J. Silva<sup>2</sup>

<sup>1</sup> Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal and SINTEF Natural Language Technologies Group, SINTEF ICT, Oslo, Norway

<sup>2</sup> Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal  
ncardoso@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt



# Overview

1. PhD motivation
2. Objectives
3. System overview
4. Experiments and results
5. Lessons learned and conclusions



# PhD motivation

- **Simple queries** work well with **simple IR** systems (term-match based document retrieval)
- What about **complex queries**?
- Current query expansion (QE) methods help...
  - More terms → matching odds increased → better retrieval results
- ... but sometimes not.
  - Bad selection of terms → drift from initial topic → noisy results



# PhD motivation (cont.)

- Most queries have *entities*, and entities have *semantic information*.
- Statistics-based QE works at *term level*. Reasoning-based QE requires working at *entity level*, where its semantic role is grounded.



Term

Entity

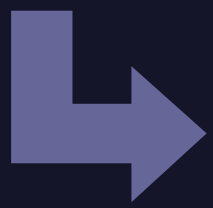
# PhD motivation (cont.)

- Why don't we try to **understand** what the user **want**, instead of **retrieving** what the user **said**?
- Why don't we **reason** to get answers instead of **guessing** terms?
- Is there a better approach for **elaborated queries**, such as queries with concrete geographic and temporal scopes?

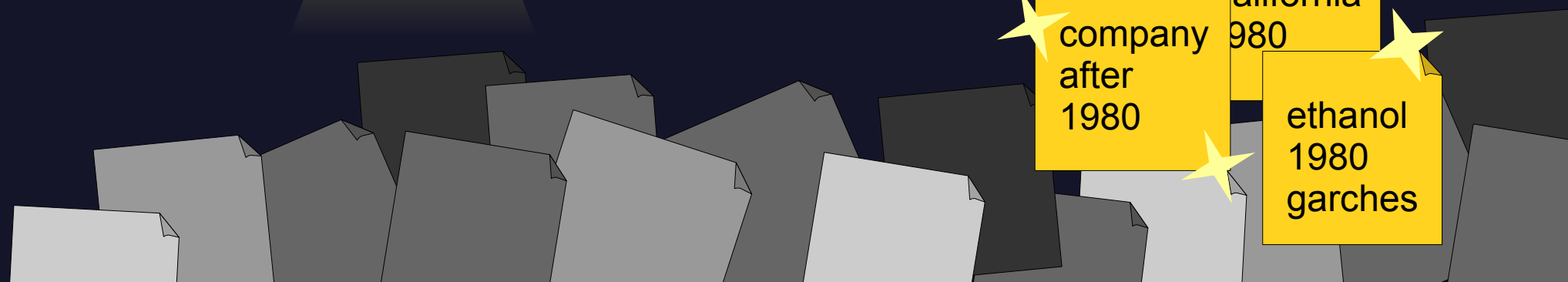
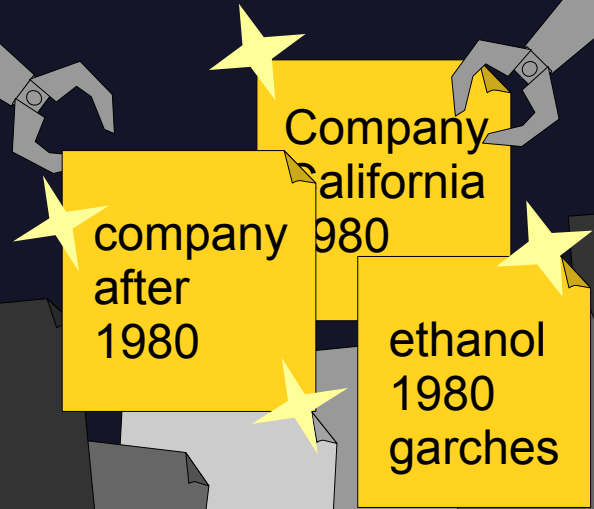
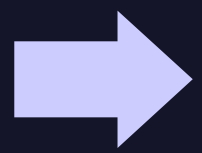
# Statistics-based Query Expansion

“Companies founded in California after 1980”

terms in cloud obtained with LucQE using the NYT collection



Query Expansion using blind relevance feedback (BRF)

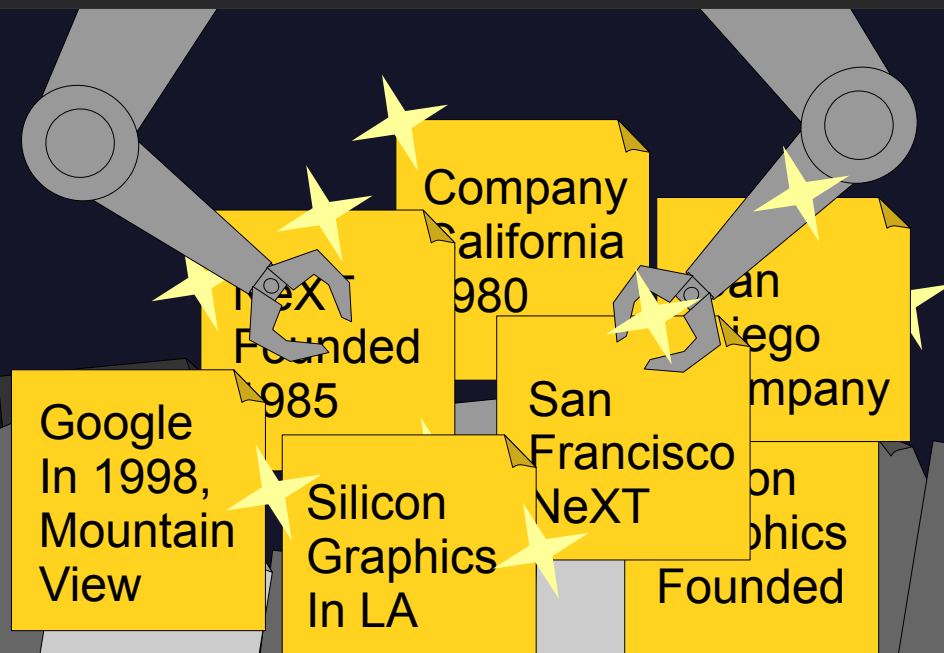


# Semantic-based Query Reformulation

“Companies founded in California after 1980”

Semantic-based Query reformulation

Entities: **California** , **1980**  
Gescope: in California  
Geographic places: **California (state)**  
Time scope: after 1980  
Timeline: [ **1980** ,...[  
Subject: <http://dbpedia.org/ontology/Company>  
Condition: formationYear, foundationPlace  
Answers: **NeXT** , **Silicon Graphics** ,...



# PhD objectives

- Build a **semantically-flavored query reformulation (SQR) approach**, using external knowledge resources and reasoning approaches to reformulate queries at entity level.
- Evaluate how suitable is a SQR approach on retrieving documents for **geographically-challenging queries**.

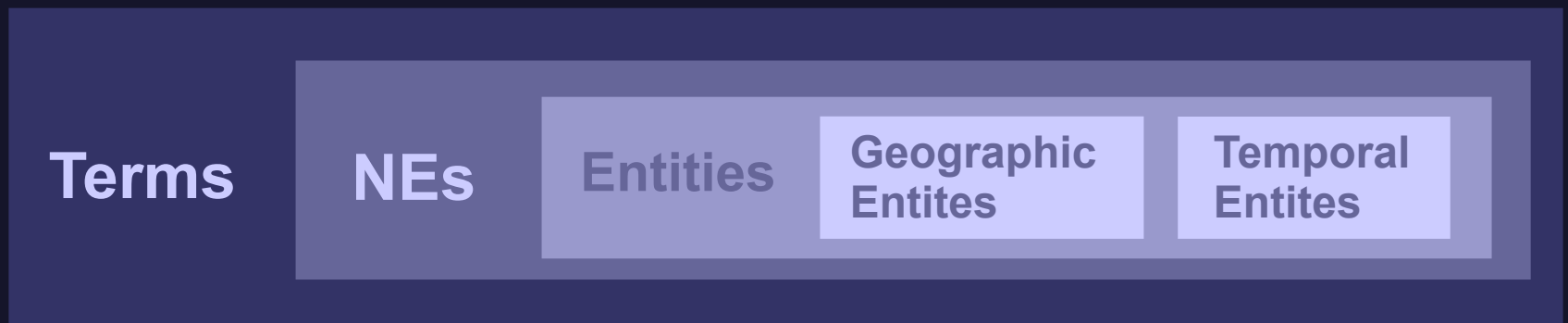
That's where NTCIR GeoTemporal task comes in...





# System overview

1. **Detect** and **ground entities** in user queries and in the *whole* document collection
  - requires a named entity recognition (NER) software.
2. Use external **knowledge bases** (Wikipedia, DBpedia, geographic ontologies) to access **more information** about entities.



# System overview

3. Index **terms** and **semantic information** (NEs, entities, places and time expressions)
4. Extend a **retrieval engine** to cope with term / semantic indexes, **reformulate queries** to use against those indexes

contents:companies contents:founded  
contents:California ne-LOCAL:'California'  
entity:California woeid:2347563 contents:after  
contents:1980 time:198\* ne-ORGANIZATION:'NeXT'  
entity:NeXT ne-ORGANIZATION:'Silicon Graphics'  
entity:Silicon\_Graphics ne-ORGANIZATION:  
'Salesforce.com' entity:Salesforce.com



# Query Parsing example

“ Where and when did Astrid Lindgren die ? ”

Question type:  
Where ,When  
Expected answer  
types: PLACE, TIME

NE: Person

Entity: [http://dbpedia.org/resource/Astrid\\_Lindgren](http://dbpedia.org/resource/Astrid_Lindgren)

rdf:label – “Astrid Lindgren”@pt

“アストリッド・リンドグレン”@jp

Property:

<http://dbpedia.org/ontology/deathPlace>

<http://dbpedia.org/ontology/deathDate>



## About: Astrid Lindgren

An Entity of Type : [person](#), from Named Graph :  
<http://dbpedia.org>, within Data Space : [dbpedia.org](#)



アストリッド・リンドグレン (Astrid Lindgren, 1907年11月14日 - 2002年1月28日) は、スウェーデンの児童書の編集者で、同時に児童文学作家でもある。彼女の著書は、世界の70ヶ国語以上に翻訳され、100以上の国で出版されている。

SPARQL query to DBpedia:

```
SELECT ?place, ?date where {  
  dbpedia:Astrid_Lindgren  
    dbpedia-owl:deathPlace?place .  
  dbpedia:Astrid_Lindgren  
    dbpedia-owl:deathDate?date .  
}
```



**Place:** <http://dbpedia.org/resource/Stockholm>

**Date:** 2002-01-28

# Query Parsing example

“japanese animators born in Tokyo”

Question type:  
none (→ Which)

Expected answer type:  
“japanese animators”

DBpedia resource:  
[http://dbpedia.org/resource/Category:Japanese\\_animators](http://dbpedia.org/resource/Category:Japanese_animators)

NE: Place  
Entity: <http://dbpedia.org/resource/Tokyo>

Property:  
<http://dbpedia.org/ontology/birthPlace>

SPARQL query to DBpedia:

```
select ?s where {  
  { ?s skos:subject  
    dbpedia:Category:Japanese_animators  
  } UNION { ?s skos:subject ?category .  
    ?category skos:broader  
    dbpedia:Category:Japanese_animators  
  }  
  ?s dbpedia-owl:birthPlace dbpedia:Tokyo .  
}
```

[http://dbpedia.org/resource/Hayao\\_Miyazaki](http://dbpedia.org/resource/Hayao_Miyazaki)  
[http://dbpedia.org/resource/Shinji\\_Higuchi](http://dbpedia.org/resource/Shinji_Higuchi)  
[http://dbpedia.org/resource/Kihachir  
%C5%8D\\_Kawamoto](http://dbpedia.org/resource/Kihachir%C5%8D_Kawamoto)

# Document retrieval example

SQR reformulated query

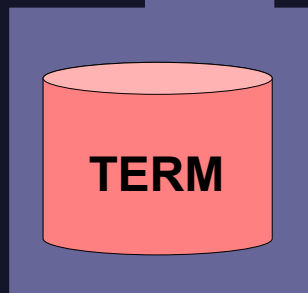
Terms

contents:where  
contents:when  
contents:'Astrid Lindgren'  
contents:die

+

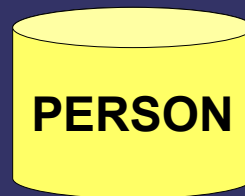
Semantic information

ne-PERSON:'Astrid Lindgren'  
entity:Astrid\_Lindgren  
ne-LOCAL:'Gutenberg'      woeid:890869  
entity:Gutenberg          time:20020128



Term index

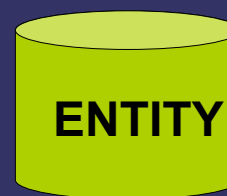
+



PERSON



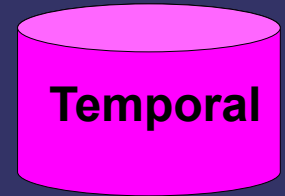
LOCAL



ENTITY



Geographic



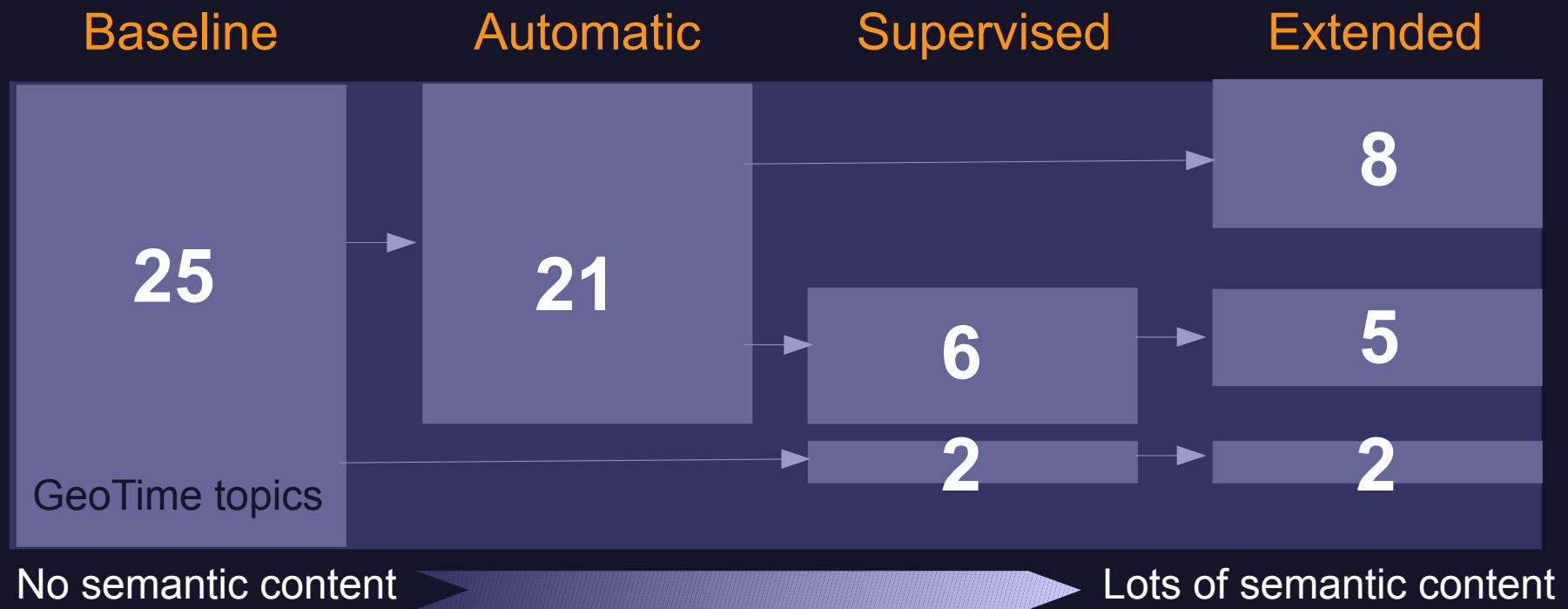
Temporal

Semantic indexes



# GeoTime experiments (EN only)

1. **Baseline** run, plain terms with no expansion
2. **Automatic** run, with DBpedia ontology lookup
3. **Supervised** run, with DBpedia ontology lookup
4. **Extended** run, with DBpedia abstract entities



Queries accumulate more semantic information from 1 to 4

# Query reformulation example

hurricane katrina make landfall united states ne-LOCAL-HUMANO-PAIS-index:"United States" woeid-index:23424977 ne-EVENT-index:"Hurricane Katrina" ne-LOCAL-FISICO-AGUAMASSA-index:"Atlantic Ocean" ne-LOCAL-HUMANO-PAIS-index:"Bahamas" ne-LOCAL-FISICO-ILHA-index:"Bahamas" ne-LOCAL-HUMANO-DIVISAO-index:Florida ne-LOCAL-HUMANO-DIVISAO-index:Louisiana ne-LOCAL-FISICO-REGIAO-index:Gulf ne-LOCAL-HUMANO-DIVISAO-index:Texas ne-LOCAL-HUMANO-DIVISAO-index:"New Orleans" woeid-index:55959709 woeid-index:23424758 woeid-index:55959686 woeid-index:2347577 woeid.index:2347602 woeid-index:615134 tg-index:20050830 tg-index:20050823

Added in Baseline run

Added in Automatic run

Added in Supervised run

Added in Extended run

# NYT 2002-2005 collectionn (EN)

Nr of <b>documents</b>	315,371
Nr of <b>NEs</b>	17,952,142
Nr of <b>classifications</b> assigned for <b>NEs</b>	18,364,572
Nr of classifications grounded to <b>entities</b>	3,344,235
Nr of classifications grounded do a <b>place</b>	588,621
Nr of docs with <b>geographic places</b>	202,624 (64%)
Nr of docs with <b>temporal expressions</b>	70,403 (22%)



# Official results

Run	mean AP
1. Baseline	0.3301
2. Automatic	<b>0.3354</b>
3. Supervised	0.3255
4. Extended	0.2978

GeoTime best:  
0.4158

- Only topic title
- No entity index at the time
- No stemming, 1:1 term:semantic index weight



# Post-hoc experiments

- Prefer entity index to NE index
- Stemming, different term:semantic index weights
- Compare/combine BRF and SQR
  1. **Baseline** run, term index, no expansion
  2. **BRF** run, term index, BRF expansion
  3. **SQR** runs, term + semantic, SQR expansion
  4. **BRF+SQR** runs, term + semantic, BRF expanded terms + SQR expanded semantic content



# Post-hoc results

MAP values (trec_eval)		no BRF	With BRF
no SQR		0.3418	0.3246
SQR	1:1	0.2869	0.2631
	2:1	0.3289	0.2958
	5:1	<b>0.3441</b>	0.3157
	10:1	0.3439	0.3269
	100:1	0.3415	0.3204
	1000:1	0.3379	0.3183

XLDB official best:  
0.3354

GeoTime best:  
0.4158



# Lessons learned

- Baselines performed well, subjects were much more important than geoscopes or timescopes
  - references to Astrid Lindgren only about her death...
- No control over term:semantic index weights → recipe for disaster
  - more semantic information means more indexes used on retrieval
  - summing partial scores from multiple indexes with BM25 unbalances retrieval focus
  - Best term:semantic ratios around 5:1



# Conclusions

- **Semantic query reformulation** can achieve good retrieval performances for geographic and temporal-flavoured queries
- **Reasoning answers** to add entities is hard, but **grounding entities** and detecting their roles is easier and very important for document ranking
- **Mixing term and semantic indexes** must be done carefully: untuned index weights may bias retrieval

# The end Questions?

## Experiments with Semantic- favored Query Reformulation of Geo-Temporal Queries

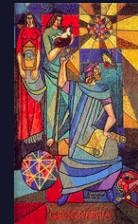
Nuno Cardoso<sup>1</sup> and Mário J. Silva<sup>2</sup>

<sup>1</sup> Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon. Portugal and SINTEF Natural Language Technologies Group, SINTEF ICT, Oslo, Norway

<sup>2</sup> Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal  
ncardoso@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt




FACULDADE DE CIÊNCIAS UNIVERSIDADE DE LISBOA




GReaSE

# 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 >>](#)

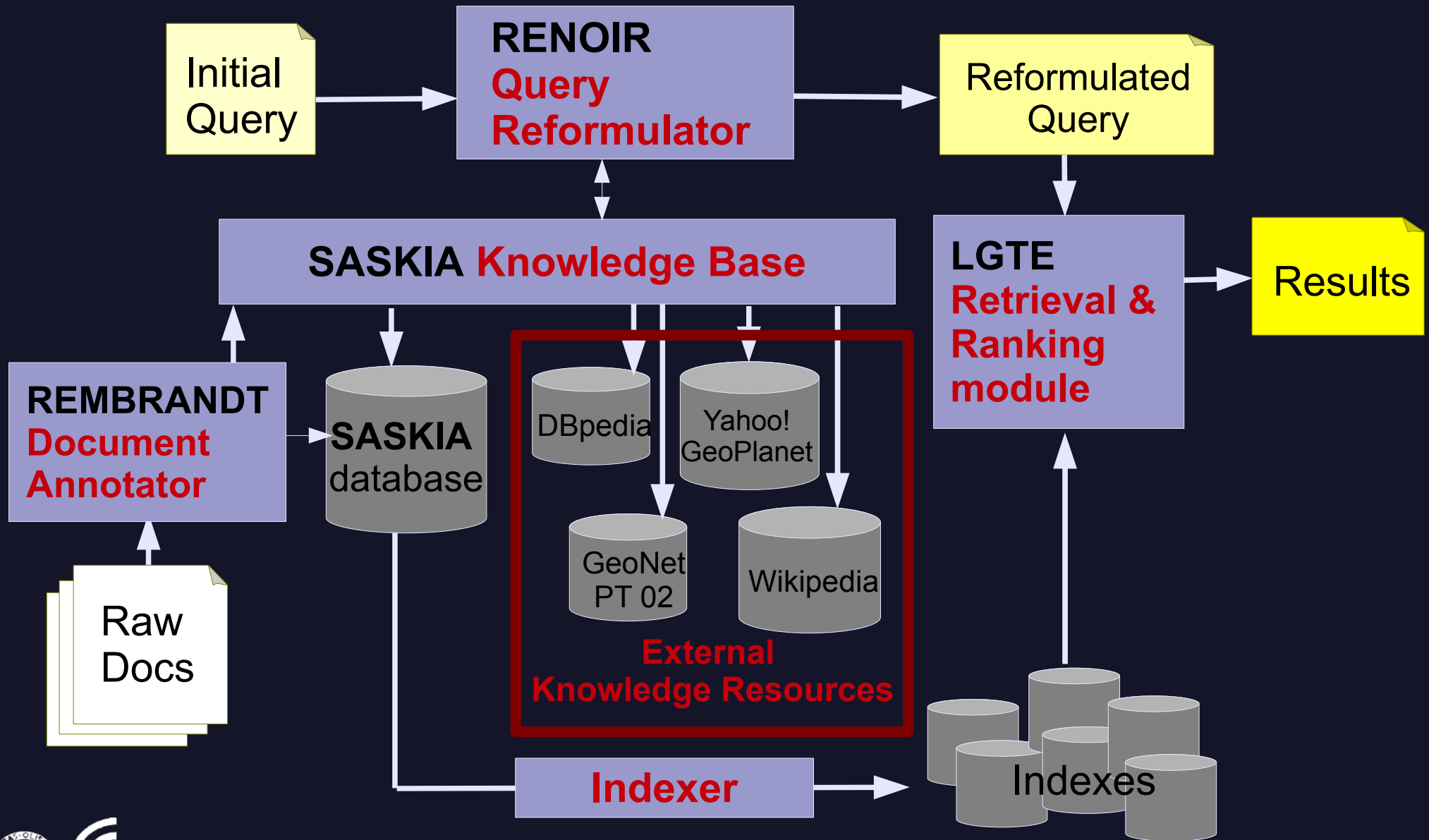
### Results

- 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 Africa. 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 navigation controls like zoom in (+), zoom out (-), and a search bar.

# GIR Architecture





# Prototype snapshots

The screenshot displays the REMBRANDT web service interface. At the top, there is a logo for REMBRANDT and a search bar with the text "forest fires in Portugal". The user is identified as "Nuno Cardoso" and the collection is "New York Times 2002-2005". The search results are displayed in a list format, with the first result titled "SPAIN SCOLDS CARELESS PUBLIC ABOUT FOREST FIRES". The text of the result is partially visible, mentioning "Spain emerges from its worst summer of forest fires in a decade, the government has begun scolding the public for the careless and sometimes malicious behavior that specialists say causes the forest fires here are starting to respect the forest fires here are starting to respect". The interface includes a sidebar with navigation options such as "Show tooltips", "Hide tooltips", "Show relations", "Hide relations", "Normal mode", "Select mode", and "Edit mode".

Português | English

Home > REMBRANDT Web Service > RENOIR QA service

forest fires in Portugal  Suggestions User: Nuno Cardoso  
Search Collection: New York Times 2002-2005

RENOIR beta Portugal x

Advanced search >>

Results 1 - 10 of 8935 for forest fires in Portugal. More info >>

Results SPAIN SCOLDS CARELE x

Show tooltips  
Hide tooltips  
Show relations  
Hide relations  
Normal mode  
Select mode  
Edit mode

? SPAIN SCOLDS CARELESS PUBLIC ABOUT FOREST FIRES

As **Spain** emerges from its worst summer of forest fires in a decade , the government has begun scolding the public for the careless and sometimes malicious behavior that specialists say causes the forest fires here are starting to respect the forest fires here are starting to respect

**Cristina Narbona**  
PESSOA INDIVIDUAL  
Sentence: 2 Term: 18  
Show relations

**Spain**  
LOCAL HUMANO PAIS  
Sentence: 4 Term: 8  
Show relations


" I am not going to apologize for saying that so **Cristina Narbona** , the minister of environment , told reporters late last month . " The ones that should apologize are the people that produce , tolerate and consent to these fires . "

About **12 23,000** forest fires had been recorded in **Spain** by last week , an increase of more than **25 percent** year , according to the Ministry of Environment .


The flames have destroyed **12 370,000** acres , an increase of **20 percent** , and killed **12 17** people

**Spain** endures the most destructive blazes in all of Europe , with nearly half the forest fires that occur each year in the **European Union** , according to a study by the **Confederation** of Workers' Commissions , a labor group that represents a large part of the government workers in the

# 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** **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): 0  
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

26