FACULDADE · DE · CIÊNCIAS   UNIVERSIDADE · DE · LISBOA

The XLDB Group at GeoCLEF 2005

Nuno Cardoso, Bruno Martins, Marcírio Chaves, Leonardo Andrade, Mário J. Silva

http://xldb.fc.ul.pt

Tumba!

- **Portuguese** Web search engine

  - ❑ Public service since 2002

  - ❑ See it in action at **http://tumba.pt**

- GREASE – Geographic Reasoning for Search Engines

- The result: GeoTumba (under development)
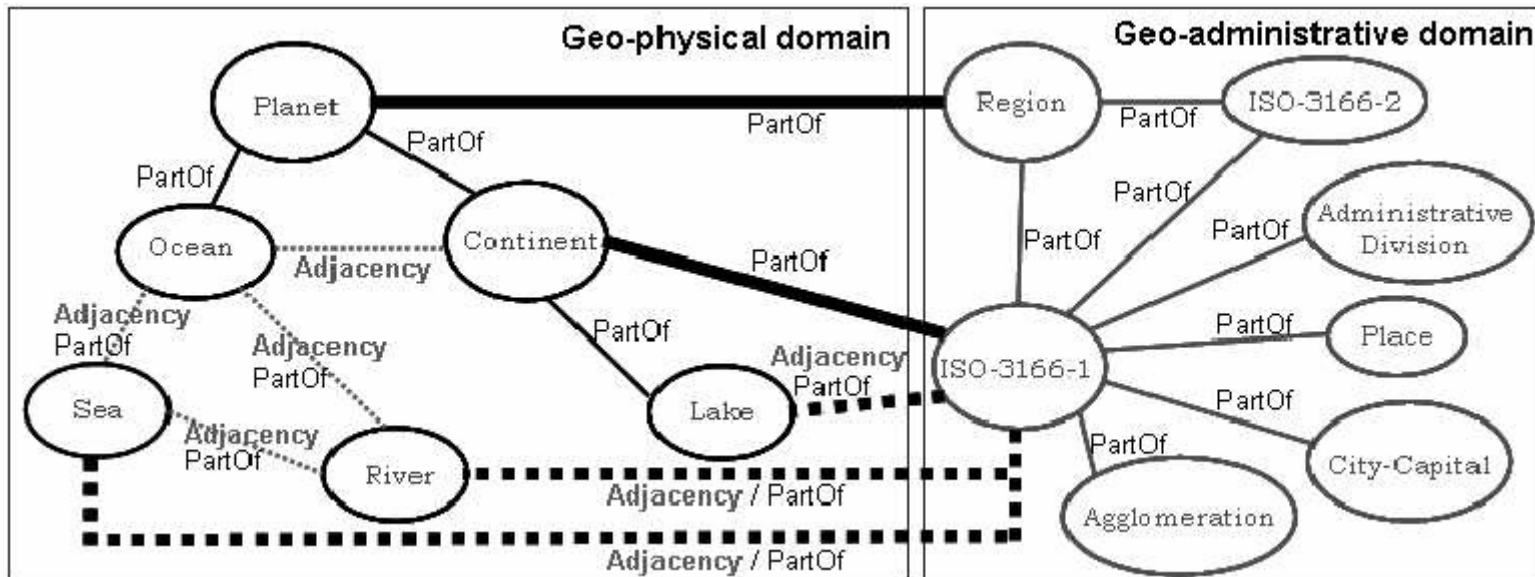
Definitions, Assumptions and Approach

- Geo-scope = footprint = focus = …

- Documents have geo-scopes

  - *One sense per discourse* assumption

- Queries have geo-scopes

- GeoIR: similarity considering undiferentiated terms + geo-scopes

Geo-IR architecture
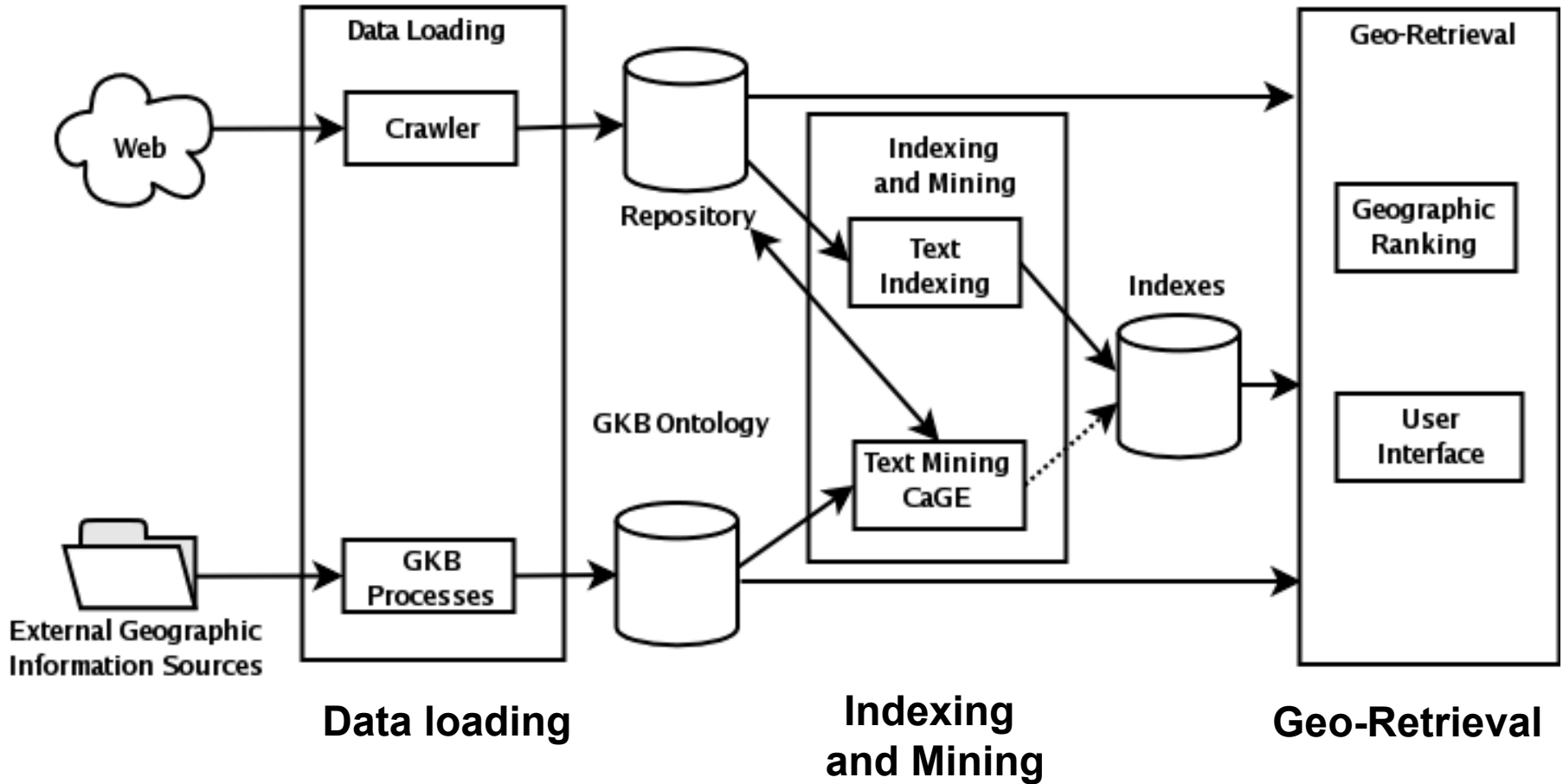
GKB – Geographic Knowledge Base

Feature types & relationships:

- Information Sources used in instance built for GeoCLEF:
  - **Wikipedia** – names of countries and capitals in four languages
  - **World Gazetteer** – cities and agglomerations with ppl > 100,000
  - **Statistics**: 12,293 features, 14,759 relationships
- TGN planned, but licensing completed only after runs submitted

Text Mining (CAGE)
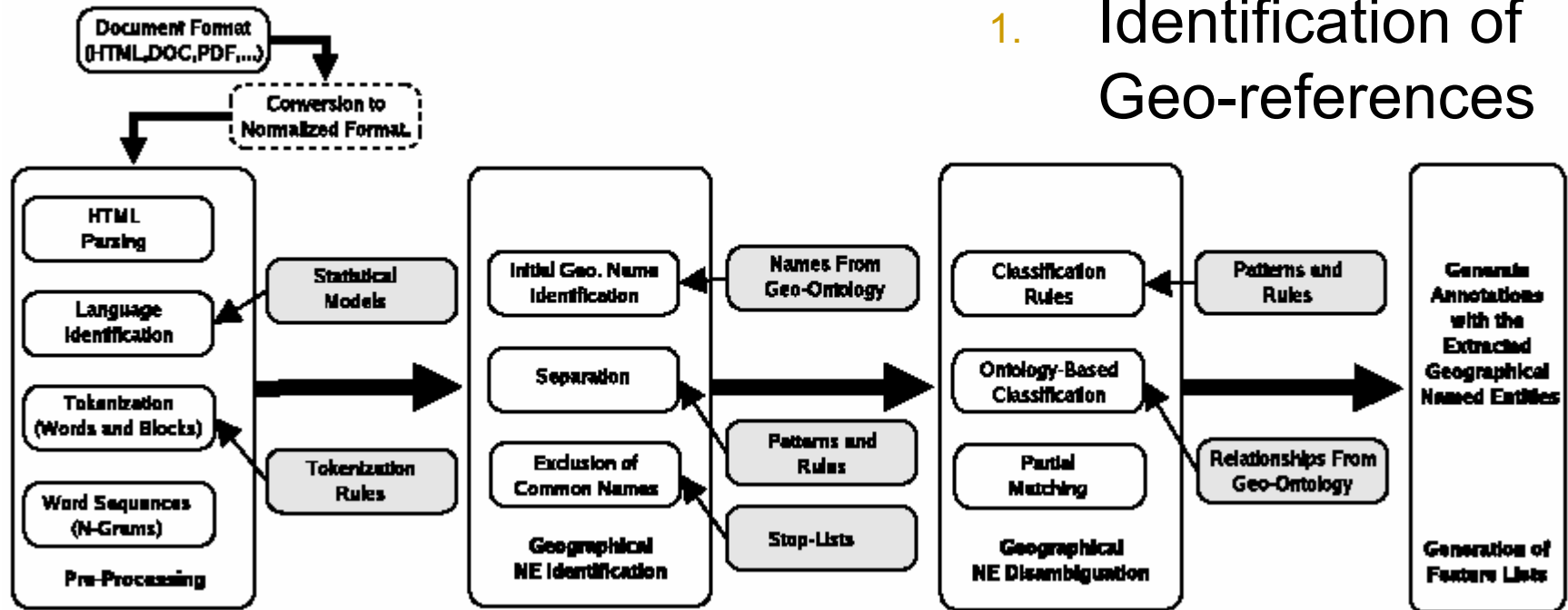
1. Identification of Geo-references
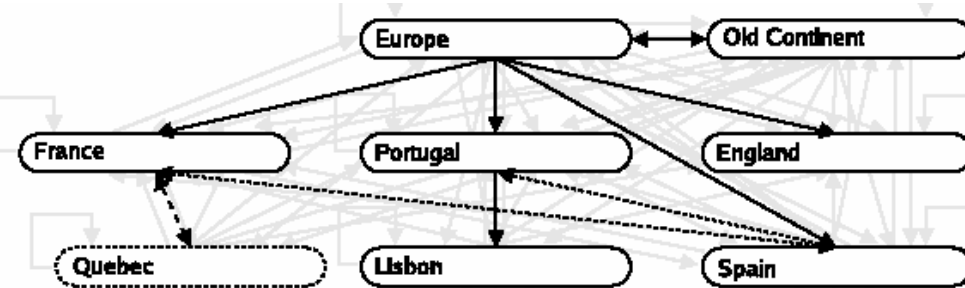
2. Scope Assignment

→ Broader/Narrower Regions   ◄······► Adjacent Regions
◄──► Equivalent Names   ◄─ ─► Related Regions

Geotumba Software Configuration used at GeoCLEF 2005

- **QuerCol**
  - Automatic Query Expansion component developed for ad hoc task
  - Transforms CLEF topics into queries

- **2 Scope assignment algorithms**
  - Most frequent geographic term
  - Graph Rank

- **Ranking by**
  - Primary: geographic similarity
  - Secondary: simplified BM25 similarity on nondifferentiated terms

Evaluation Goals

- **Scope Ranking**: compare ranking with scopes vs. use of geographic terms as non-differentiated query terms

- **Scope Assigning**: compare most frequent geographic term vs. graph algorithm

- **Location Terms Expansion**: evaluate geographic terms expansion

- **Topic Translation**: evaluate EBMT technique

XLDB@GeoCLEF2005 Monolingual EN
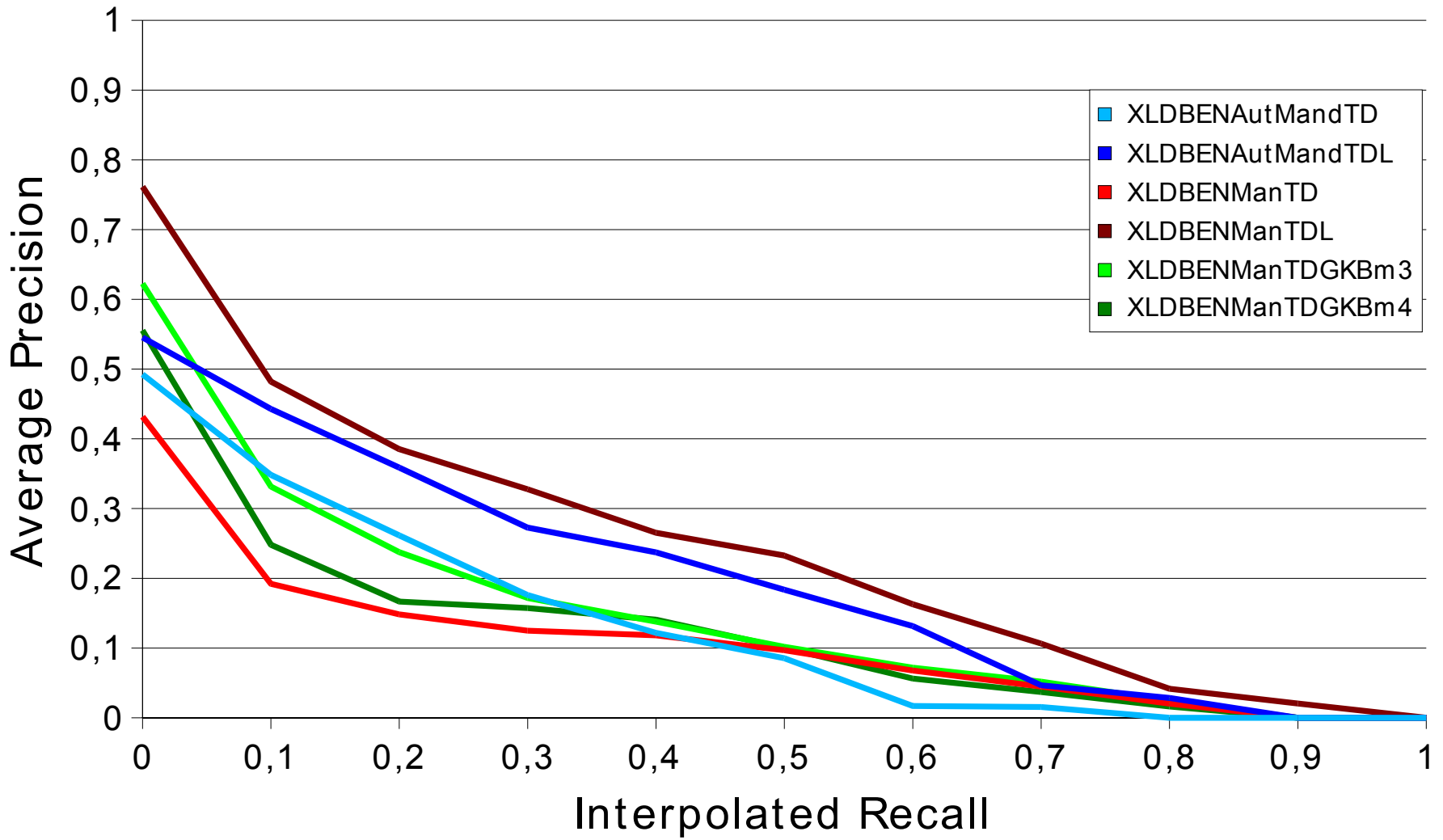
Results

- **Scope Ranking**: geo-scopes not as successful
  - ❑ We blame the assembled ontology
- **Scope Assigning**: graph algorithm always better
  - ❑ Scopes as expected, given the ontology
- **Location Terms Expansion**: manually generated queries performed better
- **Topic Translation**: monolingual runs better than bilingual

Final Observations and Future Work

- **Amount and quality of geographic knowledge have strong influence on GeoIR performance**

- **Meaning of "geographic relevance" needs clarification**

- **Future**

  - Better ranking algorithm, less sensitive to the absence of geographic terms in the ontology.

  - Evalute the effect of variation of geographic knowledge density on retrieval performance

  - Study the semantics of geographic relationships in queries.