# The XLDB Group at GeoCLEF 2005

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

Grupo XLDB - Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
{ncardoso, bmartins, mchaves, leonardo, mjs} at xldb.di.fc.ul.pt

**Abstract.** This paper describes our participation at GeoCLEF 2005. We detail the main software components of our Geo-IR system, its adaptation for GeoCLEF and the obtained results. The software architecture includes a geographic knowledge base, a text mining tool for geo-referencing documents, and a geo-ranking component. Results show that geo-ranking is heavily dependent on the information in the knowledge base and on the ranking algorithm involved.

## 1 Introduction

Over the past two years, the XLDB Group developed and operated tumba!, a search engine for the Portuguese community (`http://www.tumba.pt`) [1]. We are currently extending it to handle geographic searches, under the GREASE (Geographical REAsoning for Search Engines) project.

GREASE researches methods, algorithms and software architecture for geographical information retrieval (Geo-IR) from the web [2]. Some of the specific challenges are: 1) building geographical ontologies to assist Geo-IR; 2) extracting geographical references from text; 3) assigning geographical scopes to documents; 4) ranking documents according to geographical relevance. GeoTumba, a location-aware search engine handling *concept@location* queries, is a prototype system developed in the context of GREASE.

Our participation at GeoCLEF aimed at evaluating GeoTumba. To build a system configuration that would enable us to generate the GeoCLEF runs, we made significant adaptations to GeoTumba, including using global geographic information instead of just focusing on the Portuguese territory, and replacing the geographic ranking component (still under development) by a simpler scheme.

The rest of the paper is organized as follows: Section 2 describes GeoTumba and the software configuration that was assembled for our participation at GeoCLEF. Section 3 outlines our evaluation goals and the submitted runs. Section 4 presents an analysis on the obtained results, and finally, Section 5 draws conclusions and directions for future work.

## 2 The Geographic IR System

We take the simplistic approach of associating each document to a single scope, or none if the assignment can not be made within a certain confidence level. This is similar to the

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 997–1006, 2006.

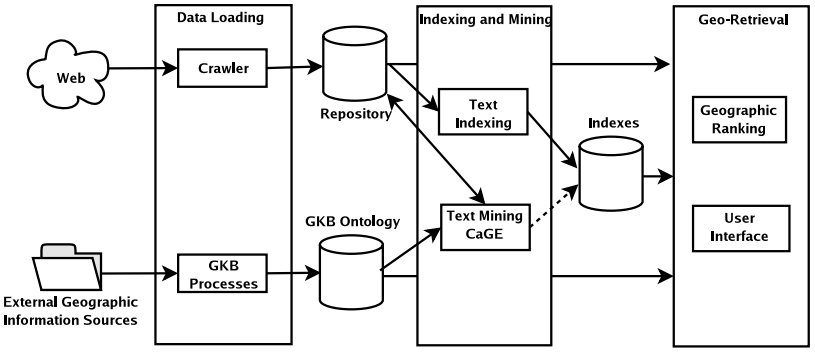


**Fig. 1.** The Geographic IR architecture

"one sense per discourse" assumption, taken in many recognition and disambiguation systems [3]. Figure 1 shows the architecture of the current Geo-IR system prototype. Information is processed in three phases:

**Data loading:** web pages are harvested into a repository by a crawling module. The geographic knowledge of GeoTumba is collected into GKB (Geographic Knowledge Base) [4]. GKB can be queried interactively to retrieve data about a geographic name or a relationship about two geographic features. It can also be used to create geographic ontologies.

**Indexing and Mining:** the geographic ontology is used by CaGE, a text mining module for recognizing geographical references and assigning documents with a corresponding geo-scope [5]. Once scopes are assigned to documents, we create indexes for fast retrieval. The indexing software of tumba! is being enhanced for indexing the geographic scopes information.

**Geo-Retrieval:** in the last phase, term indexes handle the *concept* part of the queries, while the *location* part is used as a key for fast access to documents through the scopes indexes. Result sets are generated, matching users' queries and ranked according to geographic criteria.

In the rest of this Section, we describe the main modules, GKB and CaGE, and present the software configuration that we assembled for generating the GeoCLEF submitted runs.

### 2.1 GKB – A Geographical Knowledge Base

GKB provides a common place for integrating data from multiple external sources under a common schema and exporting geographic knowledge for use by other components.

The geographical information in GKB includes names for places and other geographical features, information types (e.g. city, street, etc.), ontological relationships between the features, demographics data and geographic codes, such as postal codes.

We have developed two GKB instances: the first has detailed information about the main Portuguese territory; the second, holds information about the main regions,

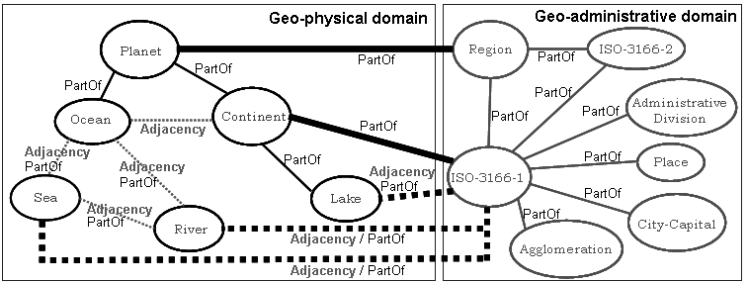


**Fig. 2.** Feature types and their relationships in the world ontology of GKB

countries, cities and places around the world in four different languages: Portuguese (PT), Spanish (ES), English (EN) and German (DE). While the first was created to support the GeoTumba service for Portugal, the latter is intended for validation of the GeoTumba software, through experiments with annotated multilingual corpora and Geo-IR evaluations covering other parts of the world, such as GeoCLEF. The geographic ontology of the world was built from two public information sources:

**Wikipedia:** on-line encyclopædia ( `http://www.wikipedia.org` ). We used its name definitions of countries and their capitals in the four supported languages. We also collected all the geo-physical names information from this source.

**World Gazetteer:** ( `http://www.world-gazetteer.com` ) information about the largest cities and agglomerations around the world. We selected those with population above 100,000.

We detail some statistics for the world ontology used in GeoCLEF elsewhere [4]. The majority of the relationships are of the `PartOf` type, while `Equivalence` and `Adjacency` relationships are much less frequent. For some types, the number of described features (number of Seas, Lakes and Regions) is much smaller than in reality because they were not available in the information sources.

Some features in GKB had to be added manually, as some GeoCLEF topics included place names like the North Sea, Caspian Sea and Siberia, which are not present on the GKB information sources.

### 2.2 CaGE – Handling Geographical References in Text

CaGE is a text mining module specifically developed to infer the geographic context from collections of documents, based on the geographic knowledge contained in a OWL ontology imported from GKB. The process of geo-referencing the textual documents is performed in two stages:

1. Identify the geographical references present in each text and weight them according to frequency.
2. Assign a corresponding geographical scope to each text, considering the geographical references, their frequency, and the relationships among them.

The geographical references are handled through a named-entity recognition (NER) procedure particularly tailored to recognizing and disambiguating geographical references over the text. Although NER is a familiar task in Information Extraction [6], handling geo-references in text presents specific challenges [7]. Besides recognizing place names, we have to normalize them in a way that specifically describes or even uniquely identifies the place in question, disambiguating them with respect to their specific type (e.g. city) and grounding them with features from the geographical ontology. CaGE follows the traditional NER architecture by combining lexical resources with shallow processing operations. It can be divided into four stages: 1) Pre-processing the documents, 2) Named-entity identification, 3) Named-entity disambiguation, and 4) Generation of feature lists [8].

After extracting geo-references, we combine the available information and disambiguate further among the different possible scopes that can be assigned to each document. Our scope assignment approach relies on a graph where the relationships between geographical concepts are specified. The geographical ontology provides the needed information. We convert it to a graph representation, weighting different semantic relationships (edges) according to their importance (i.e., equivalence relationships are more important than hierarchical relationships, which in turn are more important than adjacency relationships) and weighting different geographical concepts (nodes) according to the feature weights computed at the previous step (see Figure 3). Importance scores are then calculated for all the nodes in the graph, using a variation of the PageRank ranking algorithm [5]. After a score is computed for each feature from the ontology, we select the most probable scope for the document, by taking the highest scoring feature, or none if all features are scored below a given threshold [2].

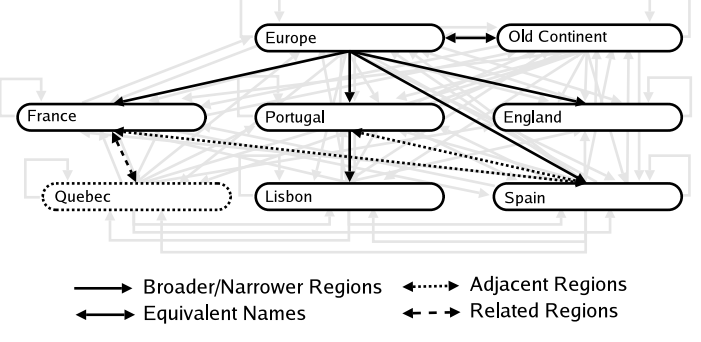


**Fig. 3.** Geographic concepts graph

### 2.3 Ranking Documents with Geo-scopes

In GeoTumba, we use geo-scopes to create new indexes supporting fast searches. The best strategies for efficiently organising this information for fast access are overviewed in [9]. We are presently pondering various similarity metrics that could be used in a global GeoTumba ranking function. As a result, we decided to participate in GeoCLEF

with a system software configuration that does not use the geographic indexes, but still ranks documents according to geographic criteria, based on the assigned scopes.

### 2.4 Software Configuration Used to Create the GeoCLEF Runs

In our GeoCLEF experiments, we used QuerCol, a query expansion component to generate queries from the CLEF supplied topics (more details about the query generation process are presented in a separate text describing our participation in the CLEF 2005 ad hoc task [10]). The changes made to GeoTumba also include:

- Replacement of the web crawler by a custom loader already used in previous evaluations, to bring the GeoCLEF newswire text collections into the repository.
- Development of a simple alternative scope assignment algorithm, that consists in simply selecting the most frequent geographical reference as the scope of a document.
- Implementation of a geo-ranking function which does not use geographic indexes. Ranking was computed in two stages: first, we ranked documents using TF $\times$ IDF weighting. Then, we ranked the given result set with a geographic similarity function. The final ranking corresponds to the set of documents ordered by a geographic rank, followed by the non-geographic rank.

The geographic similarity metric that we used in GeoCLEF is defined on a *scopes tree* extracted from the geographic concepts graph built from the geographic ontology. In this tree, we define i) $depth(X)$ as the count of edges between node $X$ and the root of the tree; ii) $ancestor(X,Y) = true$ if $X$ is on the path of $Y$ to the root node of the tree; and iii) $TD$, tree depth, the maximum $depth()$ of any node on the tree.

Given a query $Q$, a geo-scope $Scope_Q$ and a result set with documents $D_1,...,D_n$, each with a $Scope_{D_i}$ or `NULL` scope assigned, the geographic similarity $GS(Q,D_i)$ is obtained as follows:

$$GS(Q,D_i) = \begin{cases} 0 & if Scope_Q = Scope_{D_i} \\ depth(Scope_Q) - depth(Scope_{D_i}) & if\ ancestor(Scope_Q, Scope_{D_i}) = true \\ n \times TD + depth(Scope_{D_i}) - depth(Scope_Q) & if\ ancestor(Scope_{D_i}, Scope_Q) = true \\ 2 \times n \times TD & otherwise \end{cases}$$

The definition above means that the geographic similarity ranking function first ranks all the documents with the same scope as the query, then those with a narrower scope than the query, and then those with a wider scope. Finally, documents with `NULL` scopes or scopes that can not be defined as strictly narrow or wider than the scope of the query are ranked last.

## 3 Runs Description and Evaluation Goals

With our participation in GeoCLEF, we aimed at evaluating:

**Scope ranking:** measure how the ranking with the geo-scopes assigned to documents improves Geo-IR results, in comparison to including location terms in the query strings, using geographic terms as common terms, a common practice for narrowing geographic searches (e.g. '*restaurant london*') [11,12].

**Scope assigning:** when using geo-scopes, compare the graph-based algorithm against the simple scope assignment algorithm that selects the most frequent geographic entity in texts.

**Expansion of location terms:** when not using geo-scopes, measure the contribution of the expansion of geographic terms in queries to improve searches.

**Topic translation:** observe the performance of Portuguese to English bilingual runs. Our efforts were focused towards the English monolingual subtask. The bilingual runs obtained provide initial results on the performance of the machine translation system being developed by the Linguateca group at Braga, Portugal. There was no interest in creating runs derived from manual queries for this subtask.

We submitted a total of 14 runs (see Table 1). Below, we describe the creation procedures and observations intended for each of the submitted runs:

**Table 1.** The runs submitted by the XLDB group to the GeoCLEF, and their Mean Average Precision (MAP) values

| Run description | Monolingual EN | Monolingual DE | Bilingual PT->EN |
|---|---|---|---|
| (Mandatory) Automatic query generation, title + description only | XLDBENAutMandTD (MAP: 0.1183) | - | XLDBPTAutMandTD (MAP: 0.0988) |
| (Mandatory) Automatic query generation, title + description + location | XLDBENAutMandTDL (MAP: 0.1785) | - | XLDBPTAutMandTDL (MAP: 0.1645) |
| Manual query generation, title + description only | XLDBENManTD (MAP: 0.0970) | XLDBDEManTD (MAP: 0.1016) | - |
| Manual query generation, title + description + location | XLDBENManTDL (MAP: 0.2253) | XLDBDEManTDL (MAP: 0.0717) | - |
| manual query, title + description run, GKB 'PageRank'-like scopes | XLDBENManTDGKBm3 (MAP: 0.1379) | XLDBDEManTDGKBm3 (MAP: 0.1123) | XLDBPTManTDGKBm3 (MAP: 0.1395) |
| manual query, title + description run, most frequent NE scopes | XLDBENManTDGKBm4 (MAP: 0.1111) | XLDBDEManTDGKBm4 (MAP: 0.0988) | XLDBPTManTDGKBm4 (MAP: 0.1470) |

**'AutMandTD and AutMandTDL':** GeoCLEF required two fully automatic mandatory runs. The first should only use title and description information from the supplied topics, while the second should also use the location information. These two runs provide the evaluation baselines. The first indicates the performance of the non-geographical IR mechanisms being used, while the other provides the means to evaluate geographical IR against a simple baseline.

**'ManTD':** this run was generated as an intermediary step for the construction of the *ManTDL*, *TDGKBm*3 and *TDGKBm*4 runs. It provides a comparative baseline for the other submissions. We created manual queries to generate these runs, using terms from the topics's titles and descriptions, avoiding narrative terms and all related geographic terms. We did not include any location names or adjectives from the topics titles in the queries. We expanded morphologically the terms, and combined them using 'AND' and 'OR' logic operators into a single query line. As our baseline runs, the goal was to maximize recall. Precision was expected to suffer due to the lack of geographic terms on these baseline runs. These runs have a label which ends with '*ManTD*' (MANual query, Title + Description).

**'ManTDL':** We wanted to measure the efficiency of expanding and including geographical location terms in the query string, to restrict query scopes; hence, we

created these runs by inserting the scope(s) location(s) from the topic in the manual query from the '*ManTD*' runs. When the topic location scope implicitly embraces a group of countries, we extended it to the country level. For example, in the topic with the North Sea scope, the generated query string included terms like *North, Sea, England* and *Denmark*. In the case of topics with an important spatial relation (e.g. South-West of Scotland), we expanded the scope in a similar way for each location found on the narrative, like *Ayr* and *Glasgow* on the example above (notice that this was the only information used from the narratives, regarding all query strings). These runs have a label which ends with '*ManTDL*' (MANual query, Title + Description + Location).

**'TDGKBm3 and TDGKBm4':** in this run, we intended to measure the efficiency of our text mining software for assigning documents with a corresponding geographical scope, as described in Section 2. Runs labeled with '*TDGKBm*3' mark the PageRank-like scope assignment, and the labels '*TDGKBm*4' mark the most frequent geographic entity as the scope's document.

We did not submit mandatory runs for the German monolingual task, because QuerCol couldn't handle the agglutinated concepts in the topic titles properly. We found no interest in submitting these runs as the German language specificities were outside the scope of our participation in GeoCLEF.

## 4 Results

The obtained results are presented in Figures 4 and 5.Regarding the evaluation goals presented on the previous Section, we can derive from the observation of Figures 4 and 5 the following conclusions:

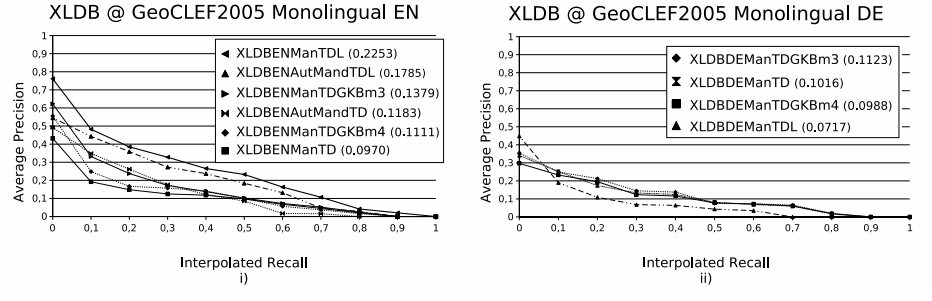


**Fig. 4.** Results of the XLDB group on the English monolingual subtask ( i) English and ii) German) of GeoCLEF 2005. In parenthesis, the MAP values of the runs.

**Scope ranking:** comparing no-scope runs vs. scope-aware runs, we observe that the runs with location terms inserted in the fully automatic query (*AutMandTDL*) ended with better precision than the runs with geographic scope ranking (*TDGKBm3* and *TDGKBm4*). We didn't expect this behaviour, as our Geo-IR is able to retrieve relevant documents to a given scope without its name on the query. A more detailed

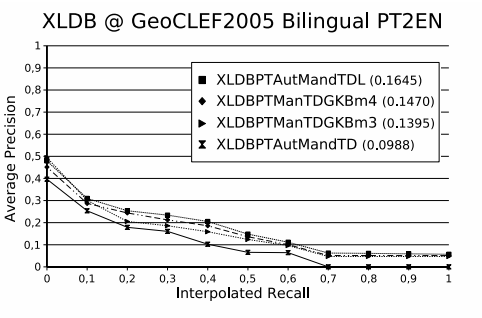


**Fig. 5.** Results of the XLDB group on the Portuguese to English bilingual subtask of GeoCLEF 2005. In parenthesis, the MAP values of the runs.

analysis of the qrels shows that this happened because both the geo-ranking method and the ontology data had limitations.

**Scope assigning:** comparing the graph-based vs. the most frequent geographical reference algorithms used to assign scopes to documents, the method based on the graph ranking algorithm (*TDGKBm3*) achieved higher precision than the alternative method of assigning the most frequent geographic reference as the document's scope (like the *TDGKBm4* runs). Analyzing the results, we can see that CaGE normally assigned the same scopes that an human would infer if he only had the same geographic knowledge passed on the world ontology.

**Expansion of location terms:** We can observe that the runs based on manual queries with expanded location terms (i.e. the *ManTDL* runs) obtained higher precision than the *AutMandTDL* runs. This reinforces our belief that relevant documents often do not explicitly contain the terms from the desired location. A Geo-IR system should consider the relationships between geographical concepts in order to retrieve relevant documents to a given location, even if they do not contain the location terms. However, the CaGE graph-ranking algorithm did not obtain better results than those used for generation of the runs based only on location names and standard text search (*AutMandTDL*). As scopes seemed to be correctly assigned, we suspect that the result was caused by lack of location names in the used ontology and a bad geographic ranking function.

**Topic translation:** The English monolingual runs exhibit better results than the bilingual runs. This results from the poor quality of the topics translation. Detailed description of these problems are included in the ad hoc participation paper [10]. This wasn't too obvious on the *ManTD* runs (they showed a similar performance), as they were created from query strings with few terms selected from the topic.

The analysis of the topic qrels shows that 61% of the relevant documents have been assigned to an unrelated or unknown scope. We realized that sub-optimal results are caused by the geographic ranking strategy adopted, and the lack of relationships in the ontology. For example, we have '*Glasgow*' as part of '*United Kingdom*', and '*United Kingdom*' as part of '*Europe*'. Yet, the record '*Scotland*' was associated

to '*United Kingdom*', and thus our geo-ranking module did not have a path from '*Glasgow*' to '*Scotland*' on the scopes tree.

Further analysis also revealed that we could have profited from using the `Adjacency` relationships on the geographic similarity metric, as we couldn't associate documents with assigned scopes like *Russia* or *Azerbaijan* to regions like *Siberia* or *Caspian Sea*.

These facts had a noticeable impact on the *TDGKBm3* and *TDGKBm4* runs, meaning that we can't make an overall evaluation of our Geo-IR, compared to the *AutMandTDL* and *ManTDL* runs, at this point.

## 5 Conclusion

For our participation in the GeoCLEF evaluation campaign, we adapted software from a geographical web search engine currently under development at our group. Our scope assignment approach is based on a two stage process, in which geographical references in the text are recognized and a geographic scope is afterwards computed for each document. A central component of the whole process is a geographical ontology, acting as the source of geographical names and relationships.

Although our scope assignment algorithm has shown to be better than a simple baseline of selecting the scopes according to the most frequent geographical references, retrieving documents using scopes was no better than the simple inclusion of the topic locations as additional terms to a standard text search. Our evaluation of the qrels has shown that the lack of information about some of the geographic concepts or their relationship to other concepts on the built ontology was the cause for poor performance in a considerable number of topics. This shows that the success of our approach strongly depends on the amount and quality of geographic knowledge that is provided to the system. However, we suspect that if too much detailed geographic information is provided, performance would also become sub-optimal.

A similar resource to GKB is the Getty Thesaurus of Geographic Names (TGN) [13]. We believe that the number of features currently in GKB is enough to assign the geographic scope to each document. We wanted to experiment this assumption with other gazetteers, and we plan to generate runs using TGN to be compared against the results obtained with GKB.

As future work, in addition to improving the scope assignment algorithm and experimenting with more comprehensive ontologies, we plan to devise and evaluate better geographic ranking functions, capable of geographically ranking documents even in the absence of geographic knowledge about terms of the query location part or in documents, and making better use of the geographic scopes.

## Acknowledgements

We thank Andreas Wichert, for the manual creation of the German queries and insight on certain aspects on diacritics expansions on German texts. Alberto Simões provided topic translations.

Thanks to all the tumba! developers and GREASE project participants. This work was financed by the Portuguese Fundação para a Ciência e Tecnologia through grant

POSI / PLP / 43931 / 2001 (Linguateca) and by grant POSI / SRI / 40193 / 2001 (GREASE). Bruno Martins is supported by FCT through grant SFRH-BD-10757-2002.

## References

1. Silva, M.J.: The Case for a Portuguese Web Search Engine. In: Proceedings of ICWI-03, the 2003 IADIS International Conference on WWW/Internet, Algarve, Portugal, IADIS (2003) 411–418
2. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P.: Adding Geographic Scopes to Web Resources. CEUS - Computers, Environment and Urban Systems, Elsevier Science (2005) In print.
3. Gale, W., Church, K., Yarowsky, D.: One sense per discourse. In: Proceedings of the 4th DARPA Speech and Natural Language Workshop. (1992)
4. Chaves, M.S., Silva, M.J., Martins, B.: GKB - Geographic Knowledge Base. Technical Report DI/FCUL TR 5-12, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (2005)
5. Martins, B., Silva, M.J.: A Graph-Based Ranking Algorithm for Geo-referencing Documents. In: Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining, Texas, USA (2005)
6. Sang, T.K., F., E., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. In Daelemans, W., Osborne, M., eds.: Proceedings of CoNLL-2003, the 7th Conference on Natural Language Learning, Edmonton, Canada (2003) 142–147
7. Martins, B., Silva, M.J.: Challenges and Resources for Evaluating Geographical IR. In: Proceedings of GIR'05, the 2nd Workshop on Geographic Information Retrieval at CIKM 2005, Bremen, Germany (2005)
8. Martins, B., Silva, M.J.: Recognizing and Disambiguating Geographical References in Web Pages. (To Appear)
9. Martins, B., Silva, M.J., Andrade, L.: Indexing and Ranking in Geo-IR Systems. In: Proceedings of GIR'05, the 2nd Workshop on Geographic Information Retrieval at CIKM 2005, Bremen, Germany (2005)
10. Cardoso, N., Andrade, L., Simões, A., Silva, M.J.: The XLDB Group participation at CLEF 2005 ad hoc task. In Peters, C., ed.: Working Notes for the CLEF 2005 Workshop, Wien, Austria (2005)
11. Kohler, J.W.: Analysing Search Engine Queries for the Use of Geographic Terms. Master's thesis, University of Sheffield (2003)
12. Martins, B., Silva, M.J.: A Statistical Study of the WPT 03 Corpus. Technical Report DI/FCUL TR-04-1, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (2004)
13. Getty Thesaurus of Geographic Names (TGN): (http://www.getty.edu/research/conducting_research/vocabularies/tgn/)