"A Method for Estimating the Precision of Place Name Matching"

by Martin Doerr[1] & Manos Papagelis[1,2]

3rd European NKOS Workshop: User-centred approaches to Networked Knowledge Organization Systems/Services (ECDL 2004)

Bath 16 September, 2004

[1]ICS-FORTH

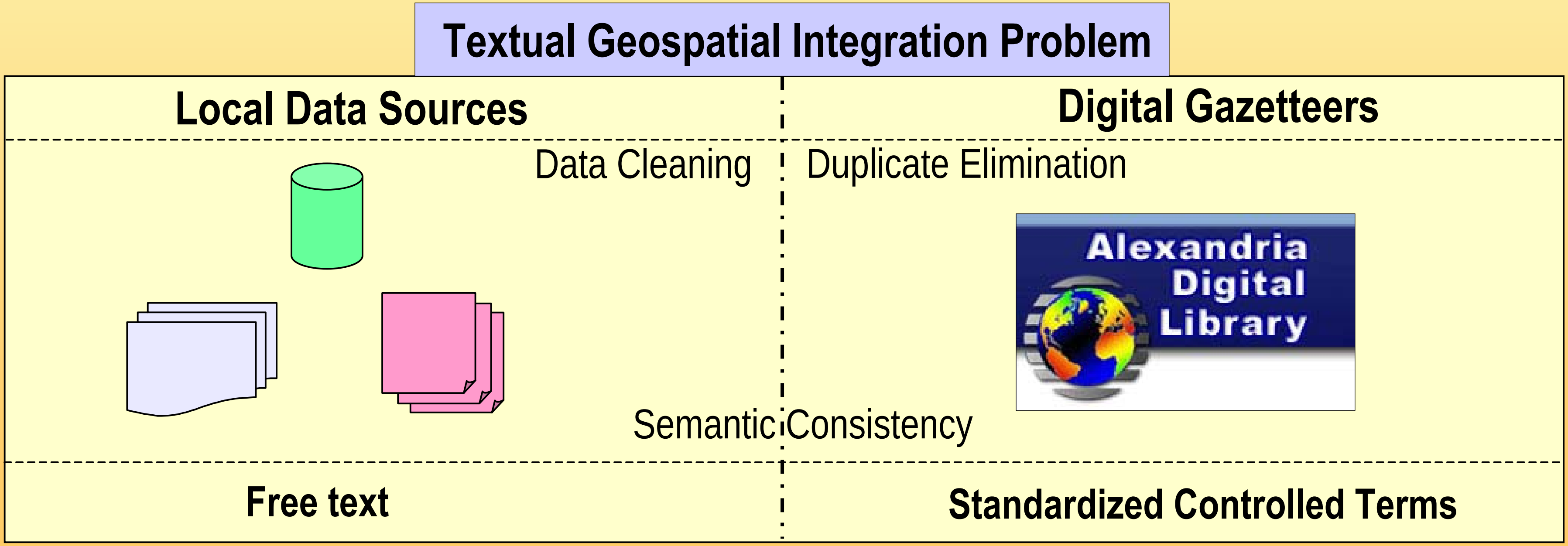
[2]Computer Science Department, University of Crete

Outline

- Problem Statement
- Objectives
- Methodology
- Experimental Evaluation and Results
- Conclusions and Discussion

Problem Statement (1/2)

- Information that resides on different data sources may be partially identical and the knowledge contained may be to some extent overlapping and complementary
- Data sources refer to locations or objects in geographic space making use of "local choice of terms"
- Digital Gazetteers systematically describe and categorize place names making use of a "global choice of terms"
- Place Name Identification: Matching of uncontrolled terms to gazetteer records



Textual Geospatial Integration Problem

Local Data Sources

Digital Gazetteers

Data Cleaning

Duplicate Elimination

Semantic Consistency

Free text

Standardized Controlled Terms

Problem Statement (2/2)

- **Identification succeeds when the geographic name :**
  - **Is found in the gazetteer**
  - **Is the only one that satisfies the place name query**
  - **Is consistent with the one intended by the data source**
- **Identification fails due to:**
  - **Non-existence**

    (Misspelling or mistyping, encoding variants or incompleteness of citation, incompleteness of digital gazetteer)
  - **Multiplicity**

    (A place name is assigned to more than one places)
  - **Semantic Inconsistency (False Positive Matches)**

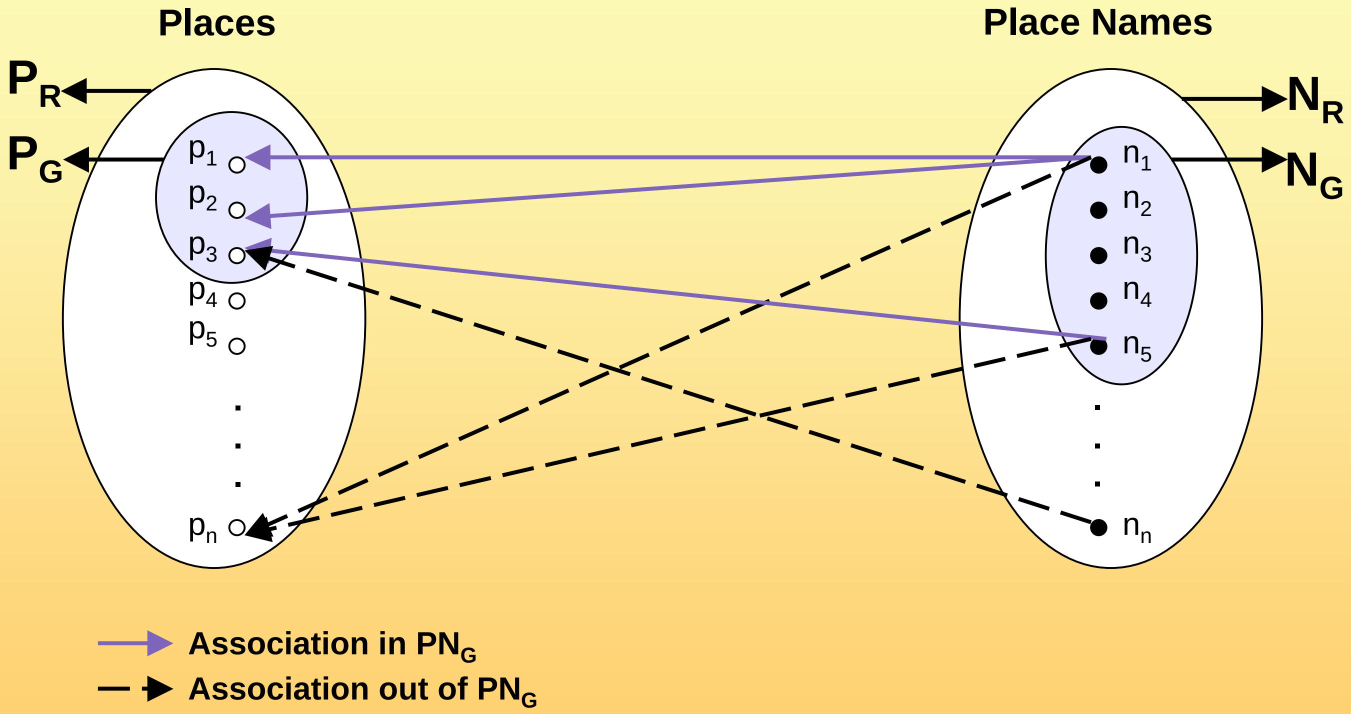
    (Mismatch between the place found and the place intended by the data source)

Objectives

- To study the cases of success and cases of failure in the place name identification process
- To provide a methodology that permits to estimate the completeness and correctness of a digital gazetteer
- In particular to estimate false matches in order:
    - To increase the automation of information integration
    - To reduce the human intervention in the process of place name identification

Methodology (1/6)

- ## Our view of the problem



→ Association in PN_G

-- → Association out of PN_G

P_R = places known in the real world

P_G = places known to the gazetteer

N_R = placenames known in the real world

N_G = placenames known to the gazetteer

Methodology (2/6)

We define as:

- $R=(P_R, N_R, PN_R)$ the real world structure, where P_R the set of real places, N_R the set of placenames and $PN_R \subset P_R x N_R$ the set of all associations between a real place and a real place name
- $G=(P_G, N_G, PN_G)$ the gazetteer structure, where P_G the set of gazetteer places, N_G the set of gazetteer placenames and $PN_G \subset P_G x N_G$ the set of all associations between a gazetteer place and a gazetteer place name
- P_{ASSOC} the probability that a place-placename association in PN_R also exists in PN_G
- F_{i_R} the global frequency of placename multiplicity *i* in R
- F_{i_G} the global frequency of placename multiplicity *i* in G

Methodology (3/6)

Assumption:

> "The process of registering a place-placename association in a gazetteer happens independently from the multiplicity of its occurrence in the real world and from the multiplicity of its occurrence in gazetteer"

Therefore:

P_{ASSOC} → constant

- If $P_{r,g}$ the probability of a placename that occurs *r* times in R to be registered *g* times in G then

$$P_{r,g} = \binom{r}{g} \times P^{g}_{ASSOC} \times (1 - P_{ASSOC})^{r-g}$$

P_{ASSOC} is constant and unknown

Methodology (4/6)

- The frequencies of a place name to be associated with zero places, or one place, or two places, … or n places in G, form the following linear equation system:

$$F_{0_G}=F_{0_R}xP_{0,0}+F_{1_R}xP_{1,0}+F_{2_R}xP_{2,0}+F_{3_R}xP_{3,0}+\ldots+F_{N_R}xP_{N,0}$$
$$F_{1_G}=F_{1_R}xP_{1,1}+F_{2_R}xP_{2,1}+F_{3_R}xP_{3,1}+\ldots+F_{N_R}xP_{N,1}$$
$$F_{2_G}=F_{2_R}xP_{2,2}+F_{3_R}xP_{3,2}+\ldots+F_{N_R}xP_{N,2}$$
$$.$$
$$.$$
$$F_{N_G}=F_{N_R}xP_{N,N}$$

➡ F_{0_G}, F_{1_G}, F_{2_G}, …, F_{N_G} values are calculated by querying the gazetteer with samples

➡ $P_{r,g}$ with depend on the one unknown probability P_{ASSOC}

Assumption:

"There are no place names without places", i.e F_{0_R} should equal to zero (!)

Therefore:

We fit the probability P_{ASSOC} until F_{0_R} becomes zero. Then, the fitting P_{ASSOC} expresses the completeness of the gazetteer for the sample.

$$P_{ASSOC} = G_{Completeness}$$

Methodology (5/6)

Sampling

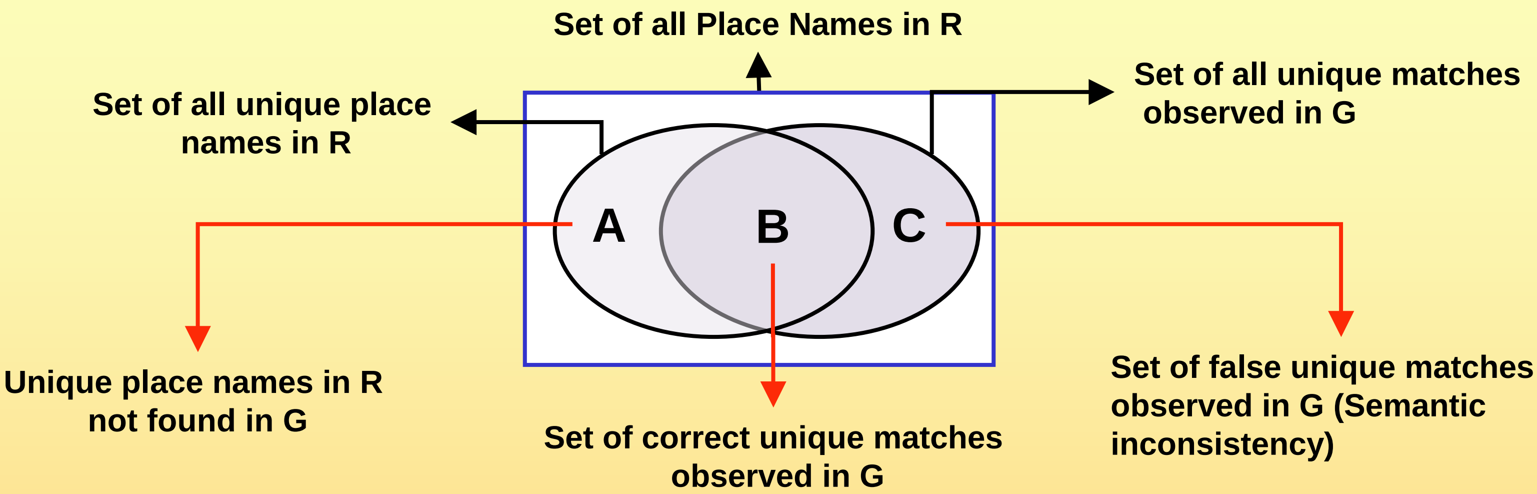
- Samples consist of place name references coming from local sources
- The sample is a surrogate for P, and the matching data for G.
- Matching problem between a sample set and a gazetteer

Assumptions

- Data sources are correct after applying data cleaning techniques
- Data sources are independent of the place-place name multiplicity of the real world structure
- Results of the matching process are related to sample, i.e the specific geographical area and application problem
- Frequencies observed from the samples should be similar to the actual frequencies F_{i_G} in the gazetteer and F_{i_R} in the real world for the target area and application.

Methodology (6/6)

- **Precision and Recall of one-to-one mappings**


Set of all Place Names in R
Set of all unique place names in R
Set of all unique matches observed in G
A
B
C
Unique place names in R not found in G
Set of correct unique matches observed in G
Set of false unique matches observed in G (Semantic inconsistency)

$$G_{\substack{\text{Semantic}\\ \text{Inconsistency}}} = \frac{F_{2_R} \times P_{2,1} + F_{3_R} \times P_{3,1} + \dots + F_{N_R} \times P_{N,1}}{F_{1_G}} \Rightarrow G_{\substack{\text{Semantic}\\ \text{Inconsistency}}} = 1 - \frac{F_{1_R} \times P_{1,1}}{F_{1_G}}$$

$$\text{Precision}_{\substack{\text{one-to-one}\\ \text{mapping}}} = \frac{\text{card}(B)}{\text{card}(B \cup C)} \Rightarrow \text{Precision}_{\substack{\text{one-to-one}\\ \text{mapping}}} = 1 - G_{\substack{\text{Semantic}\\ \text{Inconsistency}}}$$

$$\text{Recall}_{\substack{\text{one-to-one}\\ \text{mapping}}} = \frac{\text{card}(B)}{\text{card}(A \cup B)} \Rightarrow \text{Recall}_{\substack{\text{one-to-one}\\ \text{mapping}}} = \frac{F_{1_G} - G_{\substack{\text{Semantic}\\ \text{Inconsistency}}}}{F_{1_R}}$$

Experimental Evaluation and Results (1/4)

- **1000 place names originating from LUPA database**

  LUPA is a large data source that describes all known archaeological findings of stone monuments of a geographical target area (Austria), statistically well distributed from small villages to major cities.

- **Third-party authority employed:**

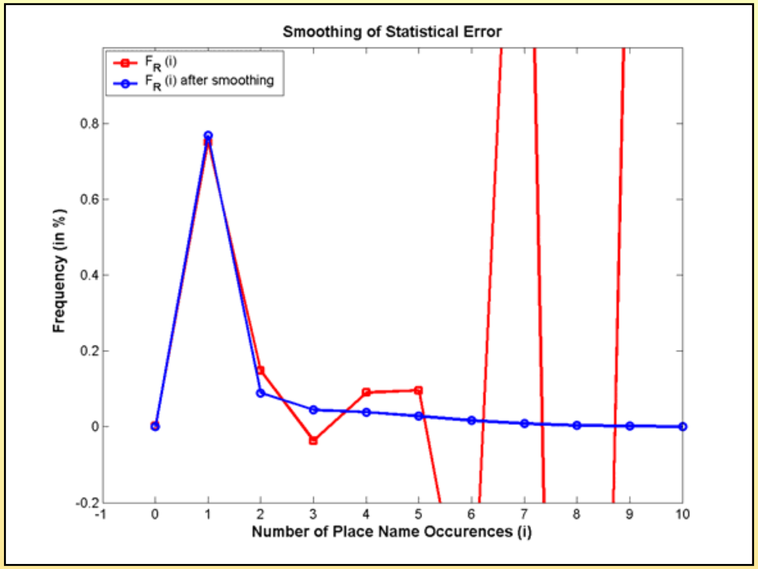
  Alexandria Digital Gazetteer (ADL) - http://www.alexandria.ucsb.edu

- **Two runs of the same sample**
  - One using "IsPartOf" relationship to narrow the searching of the place name within a specific country
  - One without using "IsPartOf" relationship, by just searching the single place name in the global scope

Experimental Evaluation and Results (2/4)

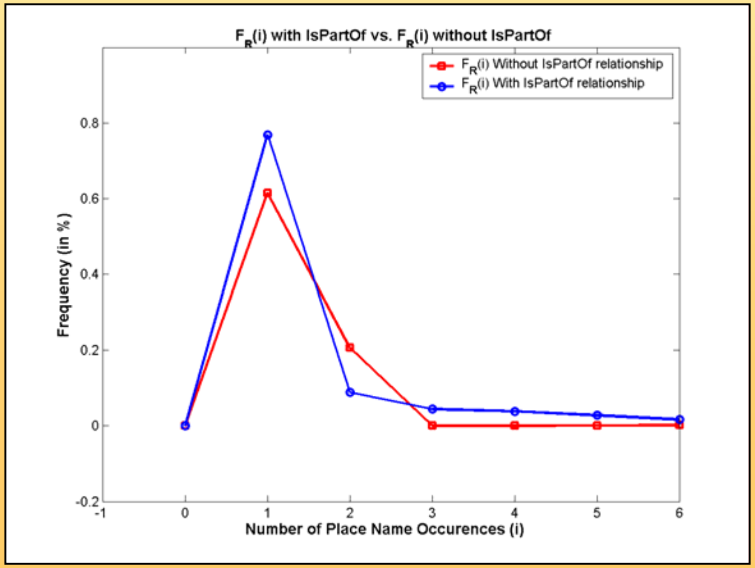
Use of a binomial function to smooth the instability (statistical error) for number of place name occurrences greater than 4 in the Gazetteer.

F_{1_R} is not affected (numerically stable)!



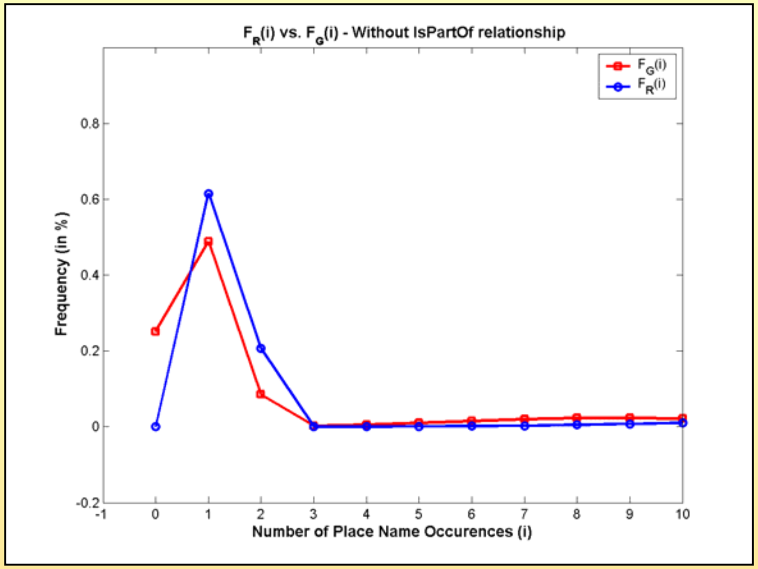
To which degree knowledge of the identity of higher levels in a hierarchy of places improves the automatic mapping process?

Answer: Look at F_{1_R} values

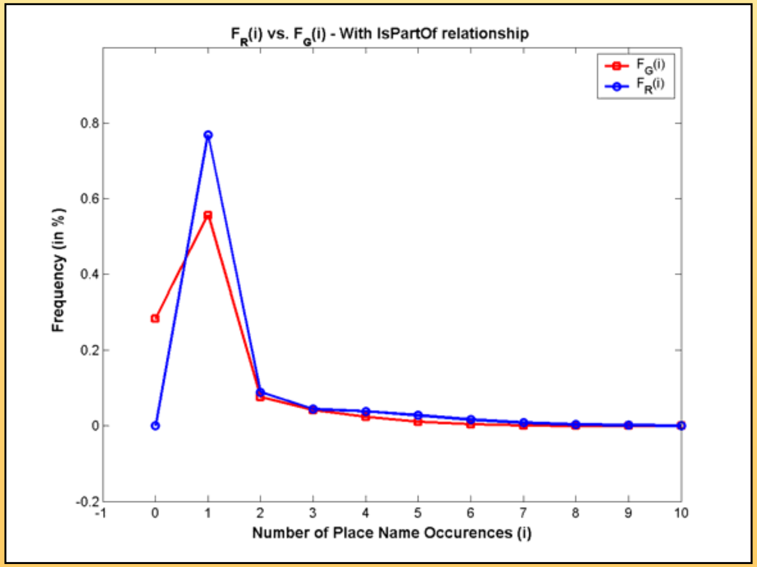


Experimental Evaluation and Results (3/4)

Frequencies of the number of place name occurrences in Gazetteer and in Real world structure when “IsPartOf” relationship is NOT applied



Frequencies of the number of place name occurrences in Gazetteer and in Real world structure when “IsPartOf” relationship is applied



Experimental Evaluation and Results (4/4)

	With IsPartOf	Without IsPartOf
$G_{\text{completeness}}$	0.6491 or 64.91%	0.6379 or 63.79%
$G_{\text{semantic\_inconsistency}}$	0.0572 or 5.72%	0.0958 or 9.58%
$\text{Precision}_{\text{one-to-one mapping}}$	0.8972 or 89.72%	0.8036 or 80.36%
$\text{Recall}_{\text{one-to-one mapping}}$	0.6490 or 64.90%	0.6374 or 63.74%

- Completeness of Gazetteer is estimated to be approximately 64%
- Less Semantic Inconsistency of one-to-one mappings when "IsPartOf" relationship is applied (from 9.57% falls to 5.72% with country knowledge)
- Precision of one-to-one mappings is estimated to be approximately 90% with IsPartOf relationship applied (~80% without "IsPartOf")
- Recall of one-to-one mappings is estimated to be approximately 64% and is not influenced by the application of "IsPartOf" relationship

Our method provides estimations with reference to a specific geographical area. However, this can be extended to any area and the Gazetteer as a whole

Conclusions (1/2)

We have presented a statistical method that permits to estimate:

- The completeness of a gazetteer with respect to a sample
- the expected precision and recall of one-to-one mappings of source place names to the gazetteer
- the semantic inconsistency that remains in one-to-one mappings
- the degree to which precision and recall are improved under knowledge of the identity of higher levels in a hierarchy of places
- It can be refined by:
  - models of the influence of placename multiplicity on registration
  - Calculating F_{0_R} by assumptions about the character of the F_{i_R} distribution i.e. determination of source misspellings.

Conclusions (2/2)

Why is this relevant:

- The method requires only the statistics of the matching process itself and no additional data. It is innovative to our knowledge. It can be easily refined.
- The semantic inconsistency is a systematic error. It can only be resolved by running manual or other heuristics on the whole data set. By knowing it, we can determine the error propagation introduced by mismatch into statistical analysis based on the matched data.
- Providing decision support for gazetteer use and development issues
- Determining the degree to which knowledge of the identity of higher levels in a hierarchy of places improves the automatic mapping process

Questions?

Thanks!