# A Method for Estimating the Precision of Place Name Matching

Martin Doerr[1] and Manos Papagelis[1, 2]

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas
P.O. Box 1385, GR-71110, Heraklion, Greece
{martin, papaggel}@ics.forth.gr
[2] Department of Computer Science, University of Crete
P.O. Box 2208, GR-71409, Heraklion, Greece

**Abstract.** Information in digital libraries and information systems frequently refers to locations or objects in geographic space. Digital gazetteers are commonly employed to identify the referred place names, in information integration and data cleaning procedures. This process fails due to missing information in the gazetteer, multiple matches or false unique matches. We study the cases of success and the cases for failure of mapping process to gazetteers. In this study, we present a statistical method that permits to estimate the completeness of a gazetteer with respect to a specific target area and application, 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 and the degree to which precision and recall are improved under knowledge of the identity of higher levels in a hierarchy of places. Our work is the first that provides estimations of completeness of a digital gazetteer and correctness of the matching process. It is useful (i) to provide decision support for gazetteer use and gazetteer development issues, (ii) to determine the error propagation introduced by mismatch into statistical analysis based on the matched data and (iii) to determine the degree to which the automatic mapping process is improved under knowledge of the identity of higher levels in a hierarchy of places. The presented method is based solely on statistical analysis of the results of the matching process itself and as such, it does not depend on any other source of additional information. This method is general and can probably be easily adopted to matching terms against other hierarchies of concepts.

**Keywords**: Information Integration, Place Name Identification, Digital Gazetteer Completeness & Correctness, Mapping of Hierarchical Concepts

## 1 Introduction

Heterogeneous data integration enables accessing a large number of data sources, developed at different times, with different organizational principles and models, satisfying different purposes and views, and supported by different platforms [1, 16]. Information that resides on different data sources may be partially *identical* and the knowledge contained is to some extent *overlapping* and *complementary*. Since infor-

mation integration [2, 3, 15] provides the prospect of querying the total of the information in a unified way, the integration process must provide an efficient way to recognize which of the referred concepts are identical [4].

On the other side, knowledge organization systems, such as glossaries, dictionaries, authority files, gazetteers, thesauri, classification and categorization systems, are employed to systematically describe, categorize and identify concepts in data cleaning procedures [12, 14] by using standardized controlled terms [17], their descriptions and semantic relationships [6, 7, 8]. These systems usually provide useful linkage between a concept and its properties. In this work, we are interested in identifying location concepts found in diverse data sources. By location concepts we mean references to locations, areas or immobile objects in the geographic space.

Digital gazetteers [6] are commonly employed to identify place names[1] in information integration procedures by matching gazetteer records, which serve as "*global choice of terms*", with words in free-text or uncontrolled data fields that come from diverse knowledge domains in databases, which serve as "*local choice of terms*". This is normally referred to as the *textual geospatial integration problem*. Identification is not always possible because the gazetteer information is hardly ever complete compared to the world structure and the source may provide insufficient information. The matching process suffers from failures due to:

- non-existence (incompleteness of a gazetteer)
- duplication (a geographic name matches to more than one place) [13]
- false positive matches (the geographic place found is not the one initially intended by the place name described in the local data source; this case results in semantic inconsistency)

In this paper we study the cases of success and the cases for failure of this mapping process. We are especially interested in estimating the precision of one-to-one mappings, the case in which one place name matches exactly one place: If those are sufficiently precise, human intervention in the matching process can be systematically reduced to the rest of the cases [11]. Based on this study, we present a statistical method that permits to estimate:

- the completeness of a gazetteer with respect to a target area
- 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

Estimations are made with respect to a specific geographical target area and application. Our method is based solely on statistical analysis of the results of the matching

[1] Actually, it remains a philosophical question how a place can be properly identified. Digital gazetteers, such as Alexandria Digital Gazetteer [9], identify places by providing geographic coordinates of it while, OpenGIS Consortium [19] accepts as true that as geographic coordinates are subject of continuous optimization they are proven insufficient to properly identify a place and they reduce the problem of identification to the unambiguous existence of the place.

process itself, of all the place names of a data source against a gazetteer and as such, it does not depend on any other source of additional information.

## 2. Study

Digital Gazetteers serve as a reference authority for geographic names providing a sound approach of assigning unique identifiers to geographic places [9, 18]. Considering the difficulties of developing, maintaining and updating a digital gazetteer, it is reasonable to assume that a gazetteer is always incomplete compared to the world structure, which is under constant evolution. Further, sources may refer to any state of geopolitical structures in the past, which are not all captured by the gazetteer. In this paragraph, we analyse the *cases of success* and the *cases for failure* in the mapping process.

### 2.1 Assumptions

For this work we assume the following:

- Digital gazetteers consist of well-defined descriptions of geographic names.
- A digital gazetteer describes a correct subset of the real world.
- Places and the non-cyclic inclusion relations between them in the real world are well-defined and well-distinguished. A maximum level of hierarchy (such as the politically independent units) can be defined. This means that well-defined levels of hierarchy can be produced.
- Geopolitical places form a hierarchical structure. This hierarchy is normally "closed" at higher levels: E.g. the politically autonomous units form a reasonably complete partitioning of the solid surface of the world. It is "open" at the lowest levels in the sense that any place might be subdivided into smaller units at some time. This is quite the opposite of what Semantic Spatial Hierarchy (SSH) model assumes for region concepts [5]. Gazetteers typically have greater coverage at the most general geographic names and less coverage at the specific geographic names.

### 2.2 Cases of success

A mapping between an undetermined geographic name used in a free text or uncontrolled data field in a data source and the controlled geographic term described in a digital gazetteer is considered successful when all of the following statements are satisfied.

- The geographic name is found in the gazetteer
- The geographic name found in the gazetteer is the only one that satisfies the query

- There is semantic consistency between the one single geographic place found in the gazetteer and the geographic place described by the free text geographic name of the data source, i.e. they mean the same area in the real world.

### 2.3 Cases for failure

A mapping between a undetermined geographic name described by free text in a data source and the controlled term geographic name described in a digital gazetteer may fail due to one of the following reasons.

#### 2.3.1 Non-existence of a geographic name

A geographic name described by free text in a data source is not found in the gazetteer. Non-existence of a geographic name occurs due to one of the following reasons.

*Misspelling or mistyping problem*

An uncontrolled geographic name described by free text in a data source is vulnerable to misspelling or mistyping problems. In this case, the mapping process fails because the misspelled geographic name is not described in the gazetteer. The case in which a misspelled geographic name is actually found in the gazetteer is regarded as failure due to semantic inconsistency.

*Encoding variants and incompleteness of citation*

Geographic names are in general "noun phrases", which means that they are composed of multiple single words, which may not all be cited, such as "Stratford upon Avon" and "Stratford". Further, there exist encoding variants (e.g. "Saint" versus "St." versus "Senkt"), as well as descriptive references of geographic names (e.g. "In the city of …", "near the tree ..").

*Incompleteness of digital gazetteer descriptions*

As gazetteers describe a part of the real world, there exist geographic names that are not, not yet or no more registered in the gazetteer. This occasion is encountered because of two reasons.

Initially, due to place description incompleteness. Gazetteers describe a subset of the world structure. Hence, there are many geographic places that are not registered in the gazetteer, mainly because of the degree of specialization that is employed by the gazetteer or because of the stage of development with respect to evolving reality. In addition, some gazetteers that describe the current state, may decide to delete obsolete places like Yugoslavia or Czechoslovakia.

Subsequently, due to incompleteness of variant names in a gazetteer. Gazetteers typically use a "variant name" property in order to assign to a geographic place alternative names. In the world structure, place names are assigned by a social group of people, for some period. The assignment depends on the group or the period, giving rise to alternative names for a place. The names themselves may change spelling over

time. Therefore, it is reasonable to assume that a digital gazetteer assigns just a part of the ever existing variant names of a geographic place description.

### 2.3.2 Duplication of a geographic name

Normally, a geographic name alone cannot identify a geographic place. This occurs because a specific geographic name may have been assigned to more than one places. In this work, we regard that the identification of the place fails, when its name is multiply used. Normally, if the wider area in which the referred place resides can be identified, the place name may be unique under this restriction. Our results indeed show this effect (see below).

### 2.3.3 Semantic Inconsistency of the mapping

Semantic inconsistency of mapping describes the case of failure in which the geographic place that is found in a gazetteer, based on a geographic name in a data source, does not correspond to the real geographic place intended by the data source.

## 3. Methodology

In this section, we present a probabilistic method that permits to compute the degree of completeness of a gazetteer, based on observation of correct mappings between the text in data sources and the controlled term geographic names described in a digital gazetteer.

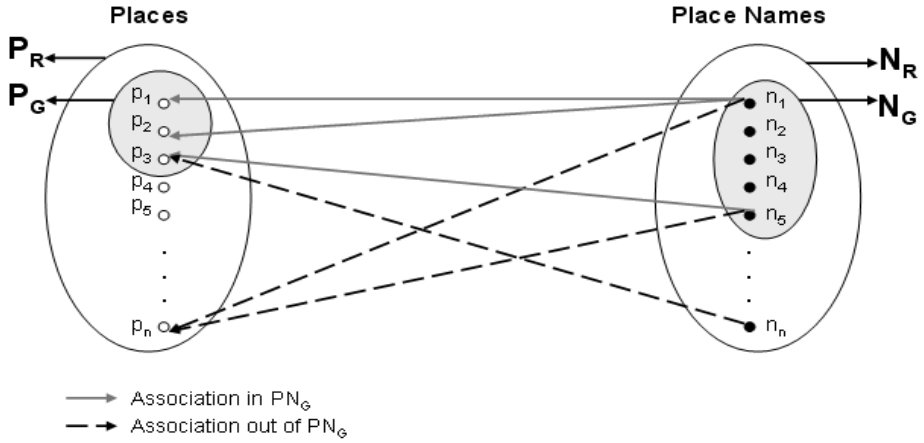


**Figure 1**. Associations between places and place names

We make the distinction between places and place names (i.e. appellations). A place can be referred to by multiple variant place names, and a place name can be assigned to one or multiple places. We describe the association between a place $p_i$

and a place name $n_j$ as a pair $(p_i, n_j)$. Figure 1 illustrates the notion of associations between a place and a place name.

If $R$ represents the real world structure $R = (P_R, N_R, PN_R)$, then we define the set of real places as $P_R$ and the set of real place names as $N_R$. We define the set of all associations between a real place and a real place name as $PN_R \subset P_R \times N_R$, where $PN_R = \{(p_i, n_j) : p_i \in P_R, n_j \in N_R\}$. Therefore, $PN_R$ represents all associations between places and place names that exist in $R$.

In analogy, if $G$ represents the gazetteer structure $G = (P_G, N_G, PN_G)$, then we define the set of places known to the gazetteer as $P_G$ and the set of place names known to the gazetteer as $N_G$. We assume that any gazetteer is incomplete, but correct: As gazetteers describe a part of the real world, we assume that $P_G \subseteq P_R$ and $N_G \subseteq N_R$. We also define the set of all associations known to the gazetteer between a gazetteer place and a gazetteer place name as $PN_G \subset P_G \times N_G$, where $PN_G = \{(p_i, n_j) : p_i \in P_G, n_j \in N_G\}$, i.e., $PN_G$ represents all associations between places and place names that exist in $G$. Accordingly, we assume that $PN_G \subseteq PN_R$, which means that the associations between places and place names in a gazetteer are a correct subset of the associations of the reality.

The notion of completeness of a Gazetteer is straightforwardly related to the percentage of place-place name associations of $PN_R$ that are found in $PN_G$:

$$G_{Completeness} = \frac{Card(PN_G)}{Card(PN_R)}$$

Before continuing, we introduce some notations. We define as:

- $P_{ASSOC}$ the probability of a place-place name association that exists in $PN_R$ also to exist in $PN_G$, i.e. the probability of a place-place name association to having been registered in Gazetteer.
- $Occ_R(n_j)$ the number of associations between places and place names that exist in $PN_R$ for a given place name $n_j$. Hence, $Occ_R(n_j) = card((p_i, n_j) : p_i \in P_R)$. This expresses the real occurrence of $n_j$ in $PN_R$.
- $Occ_G(n_j)$ the number of associations between places and place names that exist in $PN_G$ for a given place name $n_j$. Hence,

$Occ_G(n_j) = card((p_i, n_j) : p_i \in P_G)$. This expresses the occurrence of $n_j$ in the gazetteer.

- $F_{i_R}$ is the global frequency of place name multiplicity $i$ in $R$, i.e. the number of place names in $N_R$ that occur $i$ times in $PN_R$, divided by the total number of place names in $N_R$. Therefore, $F_{i_R} = \frac{card(n_j : n_j \in N_R \wedge Occ_R(n_j) = i)}{card(N_R)}$.
- $F_{i_G}$ is the global frequency of place name multiplicity $i$ in $G$, i.e. the number of place names in $N_R$ that occur $i$ times in $PN_G$, divided by the total number of place names in $N_R$. Therefore, $F_{i_G} = \frac{card(n_j : n_j \in N_G \wedge Occ_G(n_j) = i)}{card(N_R)}$.
- $\mathrm{P}_{r,g}$ the probability of a place name that occurs $r$ times in $R$ (i.e. $Occ_R(n_j) = r$) to be registered $g$ times in $G$ (i.e. $Occ_G(n_j) = g$): $P_{r,g} = \frac{card(n_j : n_j \in N_R \wedge Occ_R(n_j) = i \wedge Occ_G(n_j) = g)}{card(n_j : n_j \in N_R \wedge Occ_R(n_j) = i)}$. Because $G$ registers only true associations, it holds that $r \geq g$.

We assume that the process of registering a place-place name association happens *independently* from the *multiplicity* of its occurrence in the real world and from the *multiplicity* of its occurrence in the gazetteer. E.g. extracting associations from random texts, without systematically exploring other places or names associated with these should fulfil this assumption. As long as we do not have good reasons to doubt that there is an unbiased approach of registering place names in a gazetteer, this assumption should be a good approximation of the real probability. Under the above assumption, the probability $\mathrm{P}_{r,g}$ is equal to

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

where probability $P_{ASSOC}$ is constant for all $r, g$ and

$$P_{ASSOC} = G_{Completeness} \quad (2)$$

For example, the probability of a place name, such as “Athens”, which is associated with let us say 3 places in $R$ to be associated with 2 places in $G$ is represented as $\mathrm{P}_{3,2}$.

Then, the frequency of place names that are associated with one place in Gazetteer is given by the use of the previous definitions as:

$$F_{1_G} = F_{1_R} \cdot P_{1,1} + F_{2_R} \cdot P_{2,1} + F_{3_R} \cdot P_{3,1} + ... + F_{N_R} \cdot P_{N,1} \quad (3)$$

Putting it in plain words, the frequency of a place name that is associated with one place in $G$, is equal to the frequency of a place name that is associated with one place in $R$ multiplied by the probability to find this association registered in $G$, added to the frequency of a place name that is associated with two places in $R$ multiplied by the probability to find one of these two associations registered in $G$, added to the frequency of a place name that is associated with three places in $R$ multiplied by the probability to find one of these three associations registered in $G$, and so forth. I.e. the higher frequencies in reality contribute to the lower frequencies observed in the gazetteer due to its incompleteness.

In the same way, we compute the frequencies of a place name to be associated with one place, or two places, …, or $n$ places in G. We respectively form the following linear equation system.

$$F_{0_G} = F_{0_R} \cdot P_{0,0} + F_{1_R} \cdot P_{1,0} + F_{2_R} \cdot P_{2,0} + F_{3_R} \cdot P_{3,0} + ... + F_{N_R} \cdot P_{N,0}$$
$$F_{1_G} = F_{1_R} \cdot P_{1,1} + F_{2_R} \cdot P_{2,1} + F_{3_R} \cdot P_{3,1} + ... + F_{N_R} \cdot P_{N,1}$$
$$F_{2_G} = F_{2_R} \cdot P_{2,2} + F_{3_R} \cdot P_{3,2} + ... + F_{N_R} \cdot P_{N,2}$$
$$\vdots$$
$$F_{N_G} = F_{N_R} \cdot P_{N,1}$$

Defining

$$A = \begin{pmatrix} P_{0,0} & P_{1,0} & P_{2,0} & \cdots & P_{N,0} \\ 0 & P_{1,1} & P_{2,1} & \cdots & P_{N,1} \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & 0 & \cdots & P_{N,N} \end{pmatrix}, \vec{F}_G = \begin{pmatrix} F_{0_G} \\ F_{1_G} \\ \cdots \\ \cdots \\ F_{N_G} \end{pmatrix} \text{ and } \vec{F}_R = \begin{pmatrix} F_{0_R} \\ F_{1_R} \\ \cdots \\ \cdots \\ F_{N_R} \end{pmatrix}$$

we can find the distribution $\vec{F}_R$ of real place name occurrences by $\vec{F}_R = A^{-1} \cdot \vec{F}_G$.

The values of the matrix $A$ are all defined with reference to the unknown probability $P_{ASSOC}$. Hence, we need one more equation to solve the system. The approach we follow here is to assume, that there are no place names without places, i.e. $F_{0_R}$ should equal to zero. Under this assumption, we can fit the probability $P_{ASSOC}$ until $F_{0_R}$

becomes zero. The fitting $P_{ASSOC}$ at the time that $F_{0_R}$ falls to zero expresses the completeness of the gazetteer as given by equation (2).

### 3.1 Sampling

The method that we describe is based on sampling. Samples consist of place name references coming from local sources. We assume that these data sources are correct after applying data cleaning techniques and they are independent of the place-place name multiplicity of the real world structure. This describes a matching process between a sample set and a gazetteer. In the sequel, we are interested in how many correct and false matches we should expect. Since the place name references in the sample set refer to a specific geopolitical area, and a specific problem domain, the results of the matching process are related to the specific geopolitical target area and problem. E.g., an archaeological database we used below, covers all finds of a certain kind in Austria and Hungary. Hence, the selection is restricted to these countries, and the distribution is that of the Roman settlements rather than the modern. Therefore, there might be a slight bias towards today unimportant communities. Apart from that, we have no reasons to assume, that the frequencies observed for the sample are very different from the real frequencies $F_{i_G}$ in the gazetteer and $F_{i_R}$ in the real world for the target area (notice, that the place name multiplicity in America and Australia is generally higher than in Europe). Therefore, we approximate the frequencies $F_{i_G}$ of the gazetteer from those observed by matching against the sample. In practice, this means that our sources must be relatively "clean" of spelling errors that may produce not existing names. In particular, $F_{1_G}$ corresponds to the frequency of observed unique matches with the gazetteer (see equation (3)).

Then, we can estimate the expected recall and precision of one-to-one mappings of source place names to the Gazetteer. Figure 2 gives a picture of the problem. There is a set of one-to-one mappings that exist in $R$ and a set of one-to-one mappings that are found in $G$. However, some of the mappings registered in $G$ are due to semantic inconsistency, and some places that could be uniquely identified are not registered.

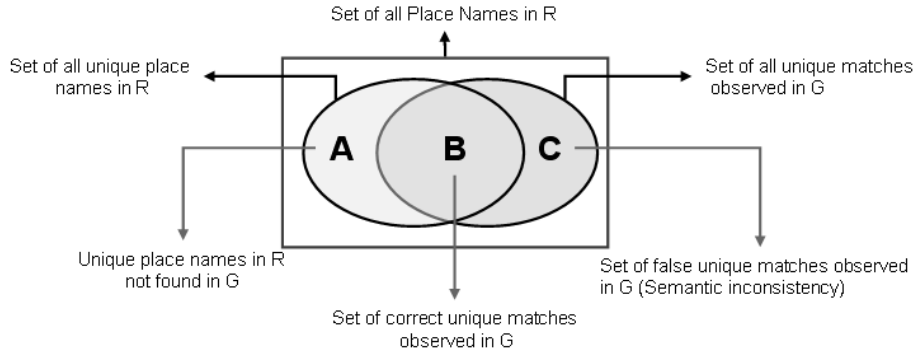


**Figure 2**. Precision and Recall of one-to-one mappings of a gazetteer.

We can determine the semantic inconsistency that may occur when identifying places using a gazetteer from the statistical analysis of a sample: When there are more than one occurrences of a place name in the real world structure but only one of them is registered in the gazetteer, then a one-to-one mapping may suffer from semantic inconsistency, if the place found in the gazetteer is different from the place we initially intend to identify. I.e. the frequency of false unique matches in the observed sample with respect to the total observed unique matches can be calculated from equation (3) as:

$$G_{\substack{Semantic \\ Inconsistence}} = \frac{F_{2_R} \cdot P_{2,1} + F_{3_R} \cdot P_{3,1} + ... + F_{N_R} \cdot P_{N,1}}{F_{1_G}} \Leftrightarrow$$

$$G_{\substack{Semantic \\ Inconsistence}} = 1 - \frac{F_{1_R} \cdot P_{1,1}}{F_{1_G}} \tag{4}$$

We compute precision as the ratio of true observed unique matches with respect to the total of observed unique matches:

$\Pr ecision_{\substack{one-to-one \\ mapping}} = \frac{card(B)}{card(B \cup C)}$, which in our context equals to

$$\Pr ecision_{\substack{one-to-one \\ mapping}} = 1 - G_{\substack{Semantic \\ Inconsistency}} \tag{5}$$

We compute recall as the ratio of true observed unique matches with respect to the total of places with a unique name in the real world:

$\mathrm{Re} call_{\substack{one-to-one \\ mapping}} = \frac{card(B)}{card(A \cup B)}$, which in our context equals to

$$\mathrm{Re} call_{\substack{one-to-one \\ mapping}} = \frac{F_{1_G} - G_{\substack{Semantic \\ Inconsistency}}}{F_{1_R}} \tag{6}$$

In the next section, we experiment with place names coming from a large data source. We compute the metrics defined above for two occasions, once by searching the place name in the global scope, and once by specializing the searching of the place name within the boundaries of its country, using an “isPartOf” relationship. Results should demonstrate, in what extent the knowledge of a higher level of representation, advances the quantity and quality of the achieved one-to-one mappings.

## 4. Evaluation of our work

### 4.1 Data set

In our experimental evaluation, we employ a set of 1000 place names originating from the LUPA database, a large data source that describes all known archaeological findings of a specific kind of a large geographical target area, statistically well distributed from small villages to major cities. The third-party authority employed is the well-known Alexandria Digital Gazetteer [9].

### 4.2 Experimental Results

We have run the sample queries twice. One time using "isPartOf" relationship to narrow the searching of the place name within a specific country, and one without using "isPartOf" relationship, by just searching the single place name in the global scope. For each place name, we compute the gazetteer occurrence for this place name, in other words, the number of associations between places and place names that exist in $G$ for the given place name. At the end of each run, this process results in computing all values of the $\vec{F}_G$ vector. We then give arbitrary values to the constant probability $P_{ASSOC}$ and compute $\vec{F}_R$ vector. We adjust $P_{ASSOC}$ until $F_{0_R}$ falls to zero. Then the solution for $\vec{F}_R$ provides an estimation of the place names occurrences in the real world structure.

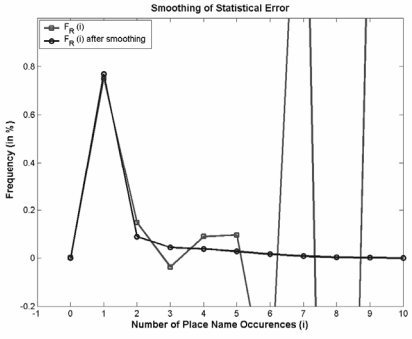


**Figure 3**. Smoothing of statistical error

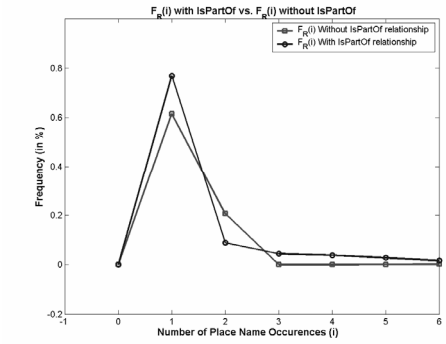


**Figure 4**. Frequency of place name occurrences in Gazetteer and in Reality

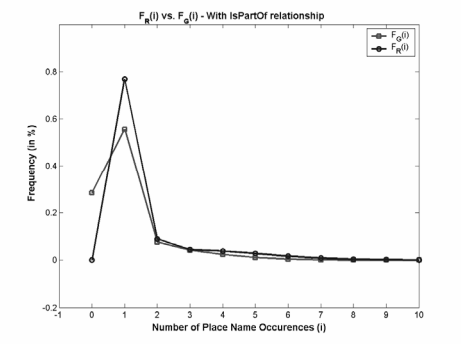


**Figure 5**. Values of $\vec{F}_R$ against values of $\vec{F}_G$ with isPartOf relationionship

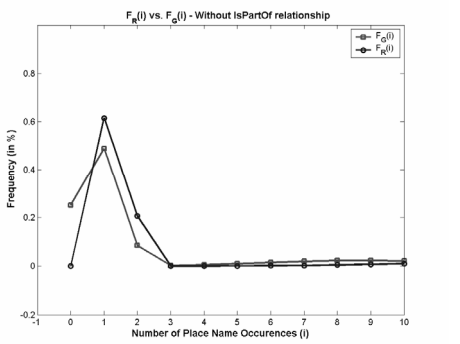


**Figure 6**. Values of $\vec{F}_R$ against values of $\vec{F}_G$ without isPartOf relationionship

For the high values of place name occurrences, we experience numerical instability due to statistical fluctuations in our limited sample. We use a binomial distribution function to smooth this instability (statistical error) for number of place name occurrences greater than 4, in order to investigate the effect of the instability on the results below. The values of $\vec{F}_R$ before and after the smoothing are depicted in Figure 3. It demonstrates that the numerical instability has no significant influence on $F_{1_R}$. Therefore, the measures in table 1 are sufficiently insensitive against this instability. Figure 4 shows the values of $\vec{F}_R$ when lacking and when having knowledge of the identity of higher levels in a hierarchy of places. Finally, Figures 5 and 6, show the calculated values of $\vec{F}_R$ against the observed values of $\vec{F}_G$ when "IsPartOf" relationship is applied and when it is not.

Based on the values of $\vec{F}_R$ vector we compute the metrics of the equations 1, 2, 3, 4 defined in the previous section for the two runs. The results are given in the Table 1:

**Table 1**. Metrics of completeness, semantic inconsistence, precision and recall of one-to-one mappings

| | **With IsPartOf** | **Without IsPartOf** |
|---|---|---|
| $G_{Completeness}$ | 0.6491 or 64.91% | 0.6379 or 63.79% |
| $G_{\substack{Semantic \\ Inconsistence}}$ | 0.0572 or 5.72% | 0.0958 or 9.58% |
| $\Pr ecision_{\substack{one-to-one \\ mapping}}$ | 0.8972 or 89.72% | 0.8036 or 80.36% |
| $\mathrm{Re}call_{\substack{one-to-one \\ mapping}}$ | 0.6490 or 64.90% | 0.6374 or 63.74% |

Results indicate that:

- *Completeness* of ADL Gazetteer is estimated to be approximately 64% for the target area and application. Knowledge of the identity of higher levels in a hierarchy of places seems to not influence the completeness of ADL Gazetteer.
- Knowledge of the identity of higher levels in a hierarchy of places results in decreasing of the one-to-one mappings that are due to semantic inconsistency. In particular, when the country of the place is known, the semantic inconsistency equals to 5.72% contrary to the case that place names are searched in the global scope, where the semantic inconsistency is 9.57%.
- Knowledge of the identity of higher levels in a hierarchy of places results in increasing of the precision of one-to-one mappings. Precision equals to 89.72% contrary to 80.36 when "IsPartOf" relationship is not used
- Recall of one-to-one mappings is about 64% and is not essentially influenced by the operation of "IsPartOf" relationship.

Our method provides estimations with reference to a specific geographical area (e.g. Austrian place names in this experiment). With suitable samples, these can be extended to any area and the gazetteer as a whole.

## 5. Conclusions & Discussion

Digital information systematically contains references to locations or objects in geographic space. Information integration processes employ digital gazetteers to efficiently identify referred locations between different sources in data cleaning procedures. However, as gazetteers are incomplete compared to the world structure and as the association between places and place names is not isomorphic, the mapping process suffers from failures due to non-existence, duplication, or semantic inconsistency. In this work, we analysed the cases of success and the cases for failure in the mapping process between a geographic name given by free text in a data source and the controlled term of the identical geographic name described in a digital gazetteer. We presented a statistical method that permits to compute the degree of completeness of a gazetteer, based on observation of mappings themselves without need for further data. We estimated the completeness of the ADL gazetteer to be about 64% for the area of Austria that our test-data cover. Our experiments show that knowledge of the wider area of a place in a hierarchy improves the precision of one-to-one mappings as much as 11,6% (from 80.36% reaches to 89.72%). In addition, the quality of the one-to-one mappings advances, by as much as 67.3%, as the semantic inconsistency of the mappings falls from 9.57% to 5.72%.

As far as our knowledge of literature permits, our work is the first that provides estimations of completeness and correctness of a digital gazetteer. We expect our method to be found functional in various upcoming studies. Suggestively we find this work significantly practical in:

- Providing decision support for gazetteer use and gazetteer development issues.

- Determining the error propagation introduced by mismatch into statistical analysis based on the matched data, such as “how many objects where found in this geographic area?”
- Defining the degree to which knowledge of the identity of higher levels in a hierarchy of places improves the automatic mapping process.

The methodology that we described is general and can probably be easily adopted to matching terms against other hierarchies of concepts, such as object types, subject terms and so forth.

## References


1. Calvanese, D., Castano, S., Guerra F., Lembo, D., Melchiori, M., Terracina, G., Ursino, D., Vincini, M.: Towards a Comprehensive Methodological Framework for Semantic Integration of Heterogeneous Data Sources. Proc. of the 8th Int. Workshop on Knowledge Representation meets Databases (KRDB 2001).
2. Wiederhold, G. (ed.) 1996. Intelligent Integration of Information. Kluwer Academic Publishers, Boston, MA.
3. Bergamaschi, S., Castano, S., De Capitani di Vimercati, S., Montanari, S., and Vincini, M. 1998. An Intelligent Approach to Information Integration. In N. Guarino (ed.) Formal Ontology in Information Systems. IOS Press.
4. Artz, J.: A Crash Course in Metaphysics for the Database Designer. Journal of Database Management. 8(4) (1997).
5. Remolina, E., Fernández, J., A., Kuipers, B., J., González, J.: Formalizing Regions in the Spatial Semantic Hierarchy: an AH-Graphs Implementation Approach. In Proc. of the Conference On Spatial Information Theory (COSIT 1999)
6. Hill, L. & Goodchild, M., 2000. Digital Gazetteer Information Exchange (DGIE). Final Report of Workshop. Available online at: http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/DGIE%20final%20report.htm
7. Hill, L. L., Zheng, Q.: Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. Knowledge: Creation, Organization and Use. Proc. of the 62nd Annual Meeting of the American Society for Information Science 1999.
8. Hill, L. L.: Core elements of digital gazetteers: placenames, categories, and footprints. Proc. of the 4th European Conference on Digital Libraries, (ECDL 2000).
9. Alexandria Digital Library Project. Available online at: http://www.alexandria.ucsb.edu
10. Dina Bitton, David J. DeWitt. “Duplicate record elimination in large data files”. ACM Transactions on Database Systems (TODS). 8(2) (1983).
11. Vassilios S. Verykios, Ahmed K. Elmagarmid b, Elias N. Houstis.Automating the approximate record-matching process. Journal of Information Sciences, 126 (1-4) (2000).
12. Mauricio A. Hernαndez, Salvatore J. Stolfo. “Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem”. Data Mining and Knowledge Discovery Journal,2(1) (1998).
13. Tak W. Yan and Hector Garcia-Molina. “Duplicate removal in information dissemination”. In Proceedings of Very Large Databases (VLDB’95).
14. Mong Li Lee, Hongjun Lu, Tok Wang Ling, and Yee Teng Ko. “Cleansing data for mining and warehousing”. In 10th International Conference on Database and Expert Systems Applications (DEXA’99)

15. Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Naveen Ashish, Pragnesh Jay Modi, Ion Muslea, Andrew G. Philpot, and Sheila Tejada. “Modeling web sources for information integration”. In Proceedings of the Fifteenth National Conference on Artificial Intelligence, (WI 1998).
16. D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. “Data integration in data warehousing”. International Journal of Cooperative Information Systems (IJCIS’00).
17. William W. Cohen. “Knowledge integration for structured information sources containing text”. In SIGIR-97 Workshop on Networked Information Retrieval (1997).
18. D. A. Smith and G. Crane. “Disambiguating geographic names in a historical digital library”. In Research and Advanced Technology for Digital Libraries: 5th European Conference (ECDL 2001).
19. OpenGIS Consortium. Available online at: http://www.opengis.org/