A Method for Estimating the Precision of Placename Matching

Martin Doerr and Manos Papagelis

Abstract—Information in digital libraries and information systems frequently refers to locations or objects in geographic space. Digital gazetteers are commonly employed to match the referred placenames with actual locations in information integration and data cleaning procedures. This process may fail due to missing information in the gazetteer, multiple matches, or false positive matches. We have analyzed the cases of success and reasons for failure of the mapping process to a gazetteer. Based on these, we present a statistical model that permits estimating 1) the completeness of a gazetteer with respect to the specific target area and application, 2) the expected precision and recall of one-to-one mappings of source placenames to the gazetteer, 3) the semantic inconsistency that remains in one-to-one mappings, and 4) the degree to which the precision and recall are improved under knowledge of the identity of higher levels in a hierarchy of places. The presented model is based on statistical analysis of the mapping process of a large set of placenames itself and does not require any other background data. The statistical model assumes that a gazetteer is populated by a stochastic process. The paper discusses how future work could take deviations from this assumption into account. The method has been applied to a real case.

Index Terms—Data mapping, knowledge and data engineering tools and techniques, database integration, data translation.

1 Introduction

Heterogeneous data integration implies 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 elements in different data sources may be partially identical and the knowledge contained may be, to some extent, overlapping or complementary. Since information integration [2], [3], [15] aims at 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]. This is not only a question of removing annoying duplicates [23], [24] but also far more a question on joining complementary units of knowledge through references to the same real-world items, which actually gives new knowledge.

Knowledge organization systems (KOS), such as glossaries, dictionaries, authority files, gazetteers, thesauri, and classification and categorization systems describe and categorize concepts and associate them with standardized identifiers or controlled terms. Besides others, they are employed to systematically detect and identify multiple representations of the same concept in data-cleaning procedures [10], [12], [14] by using their descriptions and semantic relationships [6], [7], [8] and by associating them with standardized controlled terms [17]. By concept, we understand common items in our discourse, such as real-world objects, people, places, and periods, but also established categories and even imaginary items. KOS typically provide a set of links between a concept and its names or identifiers under which it has appeared in different contexts and a set of properties and relationships that serve for recognition or disambiguation of the concept.

In this work, we are interested in the process of identifying location concepts by their names found in different data sources. By location concepts, we mean references to areas or immobile objects in the geographic space (called "feature instances" by OpenGIS [19]) by expressions that are clearly marked as reference to a location, by either tagging or being in a respective field of a data structure. Since neither names or identifiers nor description elements such as coordinates are one-to-one related to a location concept, we regard the gazetteer node ("record") that connects a set of names with a set of properties as the proper digital surrogate of the real-world location concept (note that even the preferred ID assigned by the Gazetteer may change without a change in the perception of the concept).

Digital gazetteers [6] record and describe location concepts such as geopolitical units, prominent geographic features, and sites of scientific or social interest. The primary interest is to associate each concept with its alternative names and other features that allow for its unique identification in particular geographic coordinates. Depending on the gazetteer, descriptions can be quite rich and of encyclopedic value, including historical names and facts (TGN) and detailed classification of location concepts (TGN and Alexandria). They are commonly employed to identify placenames in information integration procedures by matching gazetteer records, which serve as "*global choice of terms*," with words in

- *M. Doerr is with the Institute of Computer Science, Foundation of Research and Technology, Hellas (ICS-FORTH), Vassilika Vouton, PO Box 1385, GR 711 10, Heraklion, Crete, Greece. E-mail: martin@ics.forth.gr.*
- *M. Papagelis is with the Department of Computer Science, University of Toronto, Sandford Fleming Building, 10 King's College, M5S 3G4, Toronto, Ontario, Canada. E-mail: papaggel@cs.toronto.edu.*

Manuscript received 9 Nov. 2005; revised 4 Sept. 2006; accepted 13 Feb. 2007; published online 26 Feb. 2007.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0528-1105.
Digital Object Identifier no. 10.1109/TKDE.2007.1033.

1041-4347/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

uncontrolled data fields that come from databases or tagged sources of diverse knowledge domains, which are regarded 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

- nonexistence (incompleteness of a gazetteer),
- ambiguity (a geographic name matches with more than one place) [13], and
- false positive matches (the geographic place found is not the one intended by the placename described in the local data source; this case results in semantic inconsistency).

In this paper, we analyze the cases of success and the cases of failure of this mapping process.

We are especially interested in estimating the precision of one-to-one mappings, the case in which one placename 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 analysis, we present a statistical model that permits us to estimate

- the expected precision and recall of one-to-one mappings of source placenames to the gazetteer,
- the semantic inconsistency that remains in one-to-one mappings,
- the completeness of a gazetteer with respect to a target area, and
- the degree to which precision and recall are improved under the 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 the statistical analysis of the data produced in the matching process itself of all the placenames of a data source against a gazetteer and, as such, it does not depend on any other source of additional information. In addition, we can estimate precision and recall of the mapped set and the completeness of a gazetteer with respect to the specific target area and application. The method has been applied to a real case in which we show how precision and recall are improved when placenames are associated with information about their wider geographic area. The statistical model assumes that a gazetteer is populated by a stochastic process. The paper discusses how future work could take deviations from this assumption into account.

2 Problem Analysis

Digital Gazetteers serve as a reference authority for geographic names, providing a sound approach to 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 incomplete compared to the world structure, which is under constant evolution. Governmental authorities may have complete control over official names and properties of their geopolitical subunits until a certain level of granularity. However, sources may refer to any state of geopolitical structures in the past, which may not all be captured by the gazetteer. They may refer to arbitrarily small units. For instance, some archaeological excavation records of the city of Vienna refer to trees and walls that disappeared decades ago. Popular and local names may differ substantially from the official ones. Moreover, governmental authorities do not collaborate in creating internationally integrated gazetteers so far. In this section, we analyze the *cases of success* and the *cases for failure* in the mapping process.

2.1 Assumptions

In order to make a statistical model, we make the following, partially simplifying assumptions:

- The content of the respective data fields under analysis actually does intend to denote a place. This is the case in database schemata that foresee explicit fields to denote a place. In "geocoding" free text, parts of a text have to be recognized ("tagged") as place names [23]. This process of *named entity recognition* has its own errors, that is, taking a name for something that is not a place for a name for a place [21]. These errors depend on the recognition method and the language and are independent of the errors treated in this paper, regardless of whether the method uses a gazetteer [20]. Otherwise, our method applies to correctly tagged place names in free texts as well.
- Digital gazetteers consist of well-defined and well-distinguished descriptions of location concepts. That is, each record corresponds to precisely one concept and each concept is described by only one record in one gazetteer. (Although the first normally holds perfectly, the latter might be a question of the quality of the gazetteer).
- The location concepts in the real world are also well-defined and well-distinguished. That is, we assume that it is theoretically possible for the curators of a source and the curator of a gazetteer to communicate and to decide, for each concept referred, to in that source, if a record in the gazetteer describes a concept in that source or not. This will definitely hold for geopolitical units due to the usual legal regulations. In general, gazetteers also deal with other kinds of places, such as mountains or rivers, where the situation is more complex. This paper does not address problems of ambiguity with respect to the actual area.
- A digital gazetteer describes a correct subset of the real world.
- The process of registering a place-to-placename association happens *independently* of the *multiplicity* of its occurrence in the real world and from the *multiplicity* of its occurrence in the gazetteer. This assumption is actually the key to the statistical model presented in this paper (see (1)). It will hold, for instance, when extracting associations from random texts, *without* systematically exploring other

places or names associated with the same name. It would not hold very well if the gazetteer contained only one placename, selected to be unique—a practice not really fit for the use of a gazetteer as discussed here. In general, it is difficult to imagine other reasonable registration practices that would have a relation to the placename multiplicity. If not random, a normal selection criterion may be the size of the geographic feature or frequency of reference. (We discuss the effects if this assumption does not exactly hold in Section 4.5.)

For the study on the effect of adding information about wider geographic areas, we assume in addition that

- the noncyclic inclusion relations between location concepts in the real world are well-defined and well-distinguished. A common level of hierarchy (such as politically independent units) can be defined.

We believe that these assumptions hold for a large number of practical cases to a degree that would not substantially change our results. In this paper, we intend to demonstrate the effectiveness of our approach rather than its completeness with respect to all irregularities. Future work may extend the model to take into account all kinds of deviations from the above assumptions.

2.2 Cases of Success

A mapping between an unresolved geographic name used in an uncontrolled data field in a data source and the controlled geographic term described in a digital gazetteer is considered to be 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 referred to by the geographic name in the data source, that is, they mean the same concept in the real world.

2.3 Cases for Failure

A mapping between an unresolved geographic name referred to 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 Not Registered Geographic Names

A geographic name referred to in a data source is not found in the gazetteer. Nonexistence of a geographic name occurs due to one of the following reasons:

1. *Misspelling or mistyping problem*. An uncontrolled geographic name referred to 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 as a name of another concept is regarded as failure due to semantic inconsistency.
2. *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 (for example, "Saint" versus "St." versus "Sankt"), as well as descriptive references of geographic names (for example, "in the city of ..." and "near the tree ...").
3. *Incompleteness of digital gazetteer records*. As gazetteers describe a part of the real world, there exist geographic names that are systematically not, not yet, or no longer registered in the gazetteer. This may be due to the following:

   - *Incompleteness with respect to the concept*. Gazetteers describe a subset of the world structure. Hence, many geographic places may not be registered in the gazetteer because of the degree of specialization intended by the gazetteer, because of delays in updating changes in the evolving reality, or because of a random submission process. In addition, some gazetteers that describe only the current political situation may decide to delete obsolete places like Yugoslavia, Czechoslovakia, and South Vietnam.
   - *Incompleteness of variant names of a concept in a gazetteer*. Gazetteers typically use a "variant name" property in order to assign alternative names to a geographic place. In the world structure, placenames are assigned by a social group of people for some period. The assignment depends on the group and the period, giving rise to alternative names for a place. Speakers of different languages are just one case of such social groups. The spelling of the names may change over time. Therefore, it is reasonable to assume that a digital gazetteer assigns just a part of the total of variant names of a geographic place.

2.3.2 Duplication of a Geographic Name

Normally, a geographic name alone is not sufficient to identify a geographic place unambiguously because it may have been assigned to more than one place. This happens particularly with descriptive names (for example, Newcastle, Takayama (Japanese = high mountain) and Matsushima (Japanese = pine island)), names of saints, or emigrated communities (for example, Athens). In this work, we regard the identification of the place as failed when its name is multiply used. The actual disambiguation may then be carried out manually. If the wider area in which the referred place resides can be identified, the placename has a higher chance of being unique under this restriction. Further, one can expect that people deliberately use unambiguous placenames within local areas. Our results show the degree of this effect (see below). There may be various other advanced heuristics to disambiguate between multiply used names, but, in this paper, we are interested in modeling the principle of success and failure in the simplest

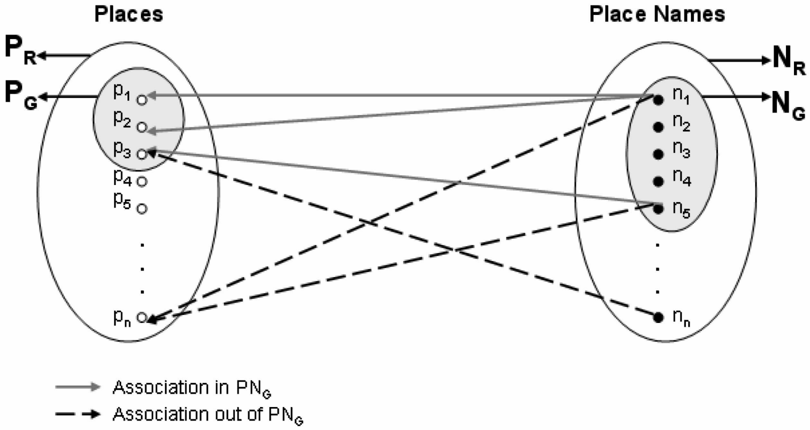

Fig. 1. Associations between places and placenames.

case first. Further, we regard it as important to devise automated procedures that differentiate results into sets with associated high precision and sets that need manual intervention so that the total manual labor to achieve an optimal result is minimized, rather than optimizing the results of the automated procedure without the possibility of reducing manual labor to improve these results.

2.3.3 Semantic Inconsistency of the Mapping

Semantic inconsistency of mapping describes the case of failure in which the geographic place that is thought to be found in a gazetteer, based on a *unique match* with the geographic name and other criteria in a data source, does not correspond to the actual geographic place intended by the data source. If there are no further heuristics available to identify such cases, only complete manual control could remedy the situation, which renders the method rather useless. In other words, the semantic inconsistency of an automated method ultimately determines its residual error, which is carried over into subsequent use of the results such as statistics based on geographic distribution.

3 Methodology

In this section, we present a statistical model that permits estimating the semantic inconsistency based on observation of correct mappings between the text in data sources and the geographic names described in a digital gazetteer, as well as other quality criteria.

We make the distinction between location concepts (from now on "places") and placenames (that is, appellations used in the real world). A place can be referred to by multiple variant placenames and a placename can be assigned to one or more places. We describe the association between a place p_i and a placename n_j as a pair (p_i, n_j). Fig. 1 illustrates the notion of associations between a place and a placename.

We define the set of all real places as P_R and the set of all real placenames as N_R. We further define the set of all associations between a real place and a real placename that exist in the real world as PN_R, where $PN_R \subset P_R \times N_R$. Based on these definitions, we declare R to represent the real-world structure $R = (P_R, N_R, PN_R)$. The three sets in the declaration of R are understood to be restricted to a certain domain of discourse, as required by an application. For instance, P_R may be just the set of all registered geopolitical subunits of a state and so forth.

In analogy, we define the set of places known to the gazetteer as P_G and the set of placenames 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 placename as PN_G, where $PN_G \subset P_G \times N_G$. Consequently, we assume that $PN_G \subseteq PN_R$, which means that the associations between places and placenames in a gazetteer are a correct subset of the associations of the reality. Based on these definitions, we declare G to represent the gazetteer structure $G = (P_G, N_G, PN_G)$.

We define the notion of the *completeness* of a Gazetteer $Comp_G$ as the percentage of place-placename associations of PN_R that are found in PN_G:

$$Comp_G = \frac{Card(PN_G)}{Card(PN_R)}.$$

In addition, we define the following:

- P_{ASSOC} is the probability of a place-placename association that exists in PN_R to also exist in PN_G, that is, the probability of a place-placename association to be registered in Gazetteer.
- $Occ_R(n_j)$ is the number of associations between places and placenames in PN_R that refer to a given placename $N_R(n_j)$. Hence,

$$Occ_R(n_j) = card((p_i, n_j) : p_i \in P_R).$$

  This expresses the real occurrence of a placename n_j in PN_R. In other words, it is the number of places one particular placename refers to. For example, three communities "Takayama" appear on the Microsoft Encarta Atlas. Therefore, $Occ_R(Takayama)$ must be greater than or equal to three.
- $Occ_G(n_j)$ is the number of associations between places and placenames in PN_G that refer to a given placename $N_G(n_j)$. Hence,

$$Occ_G(n_j) = card((p_i, n_j) : p_i \in P_G).$$

This expresses the occurrence of a placename n_j in the gazetteer. For example, two communities "Takayama" are known to the Alexandria Gazetteer. Therefore, $Occ_{AlexGaz}(Takayama) = 2$. Obviously, the maintainers of or the submitters to this Gazetteer have not searched systematically for more occurrences of Takayama (see the discussion of systematic errors).

- F_{i_R} is the global frequency of placename multiplicity i in R, that is, the number of placenames in N_R that occur i times in PN_R divided by the total number of placenames 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 placename multiplicity i in G, that is, the number of placenames in N_G that occur i times in PN_G divided by the total number of placenames in N_R. Therefore, $F_{i_G} = \frac{card(n_j : n_j \in N_G \wedge Occ_G(n_j) = i)}{card(N_R)}$.
- $\mathrm{P}_{r,g}$ is the probability of a placename n_j that occurs r times in R (that is, $Occ_R(n_j) = r$) to be registered g times in G (that is, $Occ_G(n_j) = g$):

$$P_{r,g} = \frac{card(n_j : n_j \in N_R \wedge Occ_R(n_j) = r \wedge Occ_G(n_j) = g)}{card(n_j : n_j \in N_R \wedge Occ_R(n_j) = r)}.$$

Because G registers only true associations, it holds that $r \geq g$.

Under our earlier assumption that the process of registering a place-placename association happens *independently* of the *multiplicity* of its occurrence in the real world and from the *multiplicity* of its occurrence in the gazetteer (see Section 4.5 about possible deviations), the probability $\mathrm{P}_{r,g}$ is equal to

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

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

$$P_{ASSOC} = Comp_G. \tag{2}$$

For example, the probability of the placename "Athens," which may be associated with three places in R, being associated with two places in G is represented as $\mathrm{P}_{3,2}$.

Then, the frequency of placenames that are associated with only one place in the Gazetteer is given as

$$F_{1_G} = F_{1_R} \cdot P_{1,1} + F_{2_R} \cdot P_{2,1} + P_{3_R} \cdot P_{3,1} + \ldots + F_{N_R} \cdot P_{N,1}. \tag{3}$$

To put it in words, the frequency of placenames that are associated with only one place in G is equal to the frequency of placenames that are associated with only one place in R multiplied by the probability of finding this association registered in G plus the frequency of placenames that are associated with two places in R multiplied by the probability of finding only one of these two associations in G, plus the frequency of placenames that are associated with three places in R multiplied by the probability to find only one of these three associations in G, and so forth. For example, the frequencies of placenames with higher occurrences 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 placename to be associated with one place, or two places, ..., or n places in G. We, respectively, form the following linear equation system:

$$\begin{aligned}
F_{0_G} &= F_{0_R} \cdot P_{0,0} + F_{1_R} \cdot P_{1,0} + F_{2_R} \cdot P_{2,0} + F_{3_R} \cdot P_{3,0} + \ldots + \\
&\quad F_{N_R} \cdot P_{N,0} \\
F_{1_G} &= F_{1_R} \cdot P_{1,1} + F_{2_R} \cdot P_{2,1} + F_{3_R} \cdot P_{3,1} + \ldots + F_{N_R} \cdot P_{N,1} \\
F_{2_G} &= F_{2_R} \cdot P_{2,2} + F_{3_R} \cdot P_{3,2} + \ldots + F_{N_R} \cdot P_{N,2} \\
&\vdots \\
F_{N_G} &= F_{N_R} \cdot P_{N,N}.
\end{aligned}$$

Defining

$$A = \begin{pmatrix} P_{0,0} & P_{1,0} & P_{2,0} & \ldots & P_{N,0} \\ 0 & P_{1,1} & P_{2,1} & \ldots & P_{N,1} \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ddots & \ldots \\ 0 & 0 & 0 & \ldots & P_{N,N} \end{pmatrix},$$

$$\vec{F}_G = \begin{pmatrix} F_{0_G} \\ F_{1_G} \\ \ldots \\ \ldots \\ F_{N_G} \end{pmatrix}, \text{ and } \vec{F}_R = \begin{pmatrix} F_{0_R} \\ F_{1_R} \\ \ldots \\ \ldots \\ F_{N_R} \end{pmatrix},$$

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

In other words, the function P slightly "distorts" the true distribution F_R into the distribution F_G, where each value of F_R makes a well-defined contribution to the values of F_G due to the function P. It is the typical case of an integral equation with a discrete function. If the Gazetteer was complete, A would be the identity function, and $F_G = F_R$. These kinds of mathematical problems can be easily solved if the matrix A is numerically stable, which is generally the case for diagonal-dominant triangular matrices. (Numerically stable means that small errors in the input vector do not lead to huge differences in the output vector.) It is similar to a deconvolution problem.

In our case, the values of matrix A still depend on the unknown probability P_{ASSOC}. Hence, the problem is underdetermined, that is, we have n equations for $n + 1$ variables. We need an additional equation to determine P_{ASSOC} and to solve the whole equation system. The approach we follow here is to assume that there are no placenames without places, that is, F_{0_R} should equal zero. Under this assumption, we can fit the probability P_{ASSOC} until F_{0_R} becomes zero. Fitting P_{ASSOC} at the time that F_{0_R} falls to zero expresses the completeness of the gazetteer as given by (2).

This is the basic idea of this model. It describes a kind of distorted signal, that is, a kind of convolution-deconvolution problem. In the following, we show that a reasonable approximation of the true solution can be obtained only with the data from a mapping process itself.

3.1 Sampling

In order to apply the above theory, we take the set of placename references from a local source we want to map as a representative statistical sample of the real world and we regard the frequencies observed by matching the sample with the gazetteer as a representative statistical sample of the true frequencies F_{i_G} of the complete gazetteer. Naturally, this will restrict the validity of the results to the coverage this sample naturally represents. Further, the sample should be large enough to reduce the statistical errors and independent of the place-placename multiplicity of the real-world structure.

In this paper, we present, as an application, the results from a sample of about 1,000 place references in an archaeological database which covers all finds of stone monuments in Austria and Hungary. Places are specified to the level of the closest modern community. Hence, the selection is restricted to these countries and the distribution is that of Roman settlements rather than modern. Therefore, we cannot imagine that this sample prefers a particular placename multiplicity within Austria and Hungary. Thus, we have no reason to assume that the frequencies observed for the sample are very different from the real frequencies F_{i_R} in the target area. (Notice, however, that the placename multiplicity in America and Australia is generally higher than in Europe.) However, even if they were, since we use only the application data, the parameters we compute are correct for this application in the specific case for archaeological finds in Austria and Hungary.

Further, we assume that the data sources are correct after the application of data cleaning techniques. In practice, this means that our sources must be relatively "clean" of spelling errors. Any spelling error would simulate a placename without a place and then F_{0_R} would no longer be zero (see below). The used source was well curated. Nevertheless, we could observe this effect of cleaning, as discussed in Section 4.5.

3.2 Evaluation

We are interested in how many correct and false matches we should expect in a specific matching process between a sample set and a gazetteer. In particular, F_{1_G} corresponds to the frequency of observed unique matches with the gazetteer (see (3)). From these, we can estimate the expected recall and precision of one-to-one mappings of source placenames to the Gazetteer.

As argued in Section 2, the target of this investigation is the semantic inconsistency that remains undetected. If a placename is already found twice or more in the gazetteer, a user may be able to narrow down the wider area the place falls into and thus achieve a unique match. The latter would again fall successively under this theory (that is, a new set N_G^1). Other strategies to possibly disambiguate places would need other dedicated theories. Therefore, semantic inconsistencies in the higher sets, such as placenames occurring five times to be taken for occurring twice, are not regarded here.

We can determine the semantic inconsistency that may occur when identifying places using a gazetteer from the statistical analysis of a sample: A unique mapping is inconsistent if, by chance, the one registered is not the one intended by the source. A conservative assumption is that there is a random chance 1/k of hitting the correct place (see Section 4.5 about possible deviations). That is, the frequency of false unique matches in the observed sample with respect to the total observed unique matches can be calculated as

$$G_{\substack{Semantic\\Inconsistency}} = \frac{\sum_{k=2}^{N} F_{k_R} \cdot P_{k,1} \cdot \frac{k-1}{k}}{F_{1_G}}. \tag{4}$$

To evaluate the performance of the mapping process itself, we take up the success rate, precision, and recall measures. N_R^1 is the set of placenames in N_R that occur one time in PN_R, N_G^1 is the set of placenames in N_G that occur one time in PN_G, and $G_{\substack{Semantic\\Inconsistency}}$ is the percentage of one-to-one matches between a placename in N_R^1 and a placename in N_G^1 that suffer from semantic inconsistency.

We compute *success rate* as the ratio of the correct one-to-one matches divided by the total number of places we tried to match. Correct one-to-one matches are the one-to-one matches found in the mapping process if we remove the ones that were found due to semantic inconsistency. The total number of places tried corresponds to the size of the sample. Therefore,

$$\begin{aligned}
&\text{SuccessRate}_{\substack{one-to-one\\mapping}} = \\
&\frac{card(n_j{:}n_j{\in}N_G{\wedge}Occ_G(n_j){=}1){-}card(n_j{:}n_j{\in}N_G{\wedge}Occ_G(n_j){=}1)^{*}G_{\substack{Semantic\\Inconsistency}}}{card(N_R)} \Rightarrow \\
&\text{SuccessRate}_{\substack{one-to-one\\mapping}} = \\
&\frac{\frac{card(n_j:n_j\in N_G\wedge Occ_G(n_j)=1)}{card(N_R)} - \frac{card(n_j:n_j\in N_G\wedge Occ_G(n_j)=1)^{*}G_{\substack{Semantic\\Inconsistency}}}{card(N_R)}}{\frac{card(N_R)}{card(N_R)}} \Rightarrow \\
&\text{SuccessRate}_{\substack{one-to-one\\mapping}} = F_{1_G} - F_{1_G} \cdot G_{\substack{Semantic\\Inconsistency}}.
\end{aligned} \tag{5}$$

In information retrieval, precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. In our context, we compute precision as the ratio of the correct one-to-one matches with respect to all one-to-one matches found in the mapping process:

$$\begin{aligned}
&\text{Precision}_{\substack{one-to-one\\mapping}} = \\
&\frac{card(n_j{:}n_j{\in}N_G{\wedge}Occ_G(n_j){=}1){-}card(n_j{:}n_j{\in}N_G{\wedge}Occ_G(n_j){=}1)^{*}G_{\substack{Semantic\\Inconsistency}}}{card(n_j:n_j\in N_G\wedge Occ_G(n_j)=1)} \Rightarrow \\
&\text{Precision}_{\substack{one-to-one\\mapping}} = \\
&\frac{\frac{card(n_j:n_j\in N_G\wedge Occ_G(n_j)=1)}{card(N_R)} - \frac{card(n_j:n_j\in N_G\wedge Occ_G(n_j)=1)^{*}G_{\substack{Semantic\\Inconsistency}}}{card(N_R)}}{\frac{card(n_j:n_j\in N_G\wedge Occ_G(n_j)=1)}{card(N_R)}} \Rightarrow \\
&\text{Precision}_{\substack{one-to-one\\mapping}} = \frac{F_{1_G} - F_{1_G} \cdot G_{\substack{Semantic\\Inconsistency}}}{F_{1_G}} \Rightarrow \\
&\text{Precision}_{\substack{one-to-one\\mapping}} = 1 - G_{\substack{Semantic\\Inconsistency}}.
\end{aligned} \tag{6}$$

In information retrieval, recall is defined as the ratio of the number of relevant records retrieved to the total number of relevant records in the database. In our context, we define recall as the ratio of the correct one-to-one matches with respect to the total number of the placenames we tried to match and are registered in the gazetteer. By the definition, recall is inversely proportional to the completeness of the gazetteer or, otherwise, $Recall \sim \frac{1}{Comp_G}$. Intuitively, if a low success rate is due to low completeness, recall is high. If success rate and completeness are 100 percent, recall is one. The success rate cannot exceed the completeness, that is, recall is $\leq$ 1. We compute recall as the ratio of the correct one-to-one matches divided by the total number of placenames that tried to match (that is, success rate) with respect to the completeness of the gazetteer, as is computed by our method. Therefore,

$$\begin{aligned}
&\mathrm{Recall}_{\substack{one-to-one\\mapping}} = \\
&\frac{\frac{card(n_j{:}n_j\in N_G\wedge Occ_G(n_j)=1)-card(n_j{:}n_j\in N_G\wedge Occ_G(n_j)=1)^{*}G_{\substack{Semantic\\Inconsistency}}}{card(N_R)}}{Comp_G} \Rightarrow \\
&\mathrm{Recall}_{\substack{one-to-one\\mapping}} = \frac{SuccessRate}{Comp_G} \Rightarrow \\
&\mathrm{Recall}_{\substack{one-to-one\\mapping}} = \frac{F_{1_G} - F_{1_G} \cdot G_{\substack{Semantic\\Inconsistency}}}{Comp_G}.
\end{aligned} \tag{7}$$

In Section 4, we experiment with placenames coming from a larger data source. We compute the metrics defined above for two occasions, once by searching the placename in the global scope and once by specializing the searching of the placename within the boundaries of its country using a "part of" relationship. The results demonstrate to what extent knowledge of a higher level of representation improves the quantity and quality of the achieved one-to-one mappings.

4 Application

4.1 Data Set

In our experimental evaluation, we employ a set of around 1,000 placenames originating from the "ubi erat lupa" database (http://www.ubi-erat-lupa.org), a large data source that describes all known archaeological findings of Roman stone monuments of a large geographical target area, statistically well-distributed from small villages to major cities. The placenames are those of the current community in which the object was found and its wider geopolitical units. The database is well maintained by specialists.

4.2 Experimental Scenario

The third-party authority employed in the placename matching process is the well-known Alexandria Digital Gazetteer [9]. In particular, we exploit one of the independent stateless services of the Alexandria Digital Library (ADL) Gazetteer Protocol v1.2, defined as

$$\text{reports} \leftarrow \text{query}(\text{query}, \{\,''\text{standard}\,''|\,''\text{extended}\,''\}\\ [, \text{geometry language}]).$$

The service returns reports with gazetteer entries that satisfy a *query*. Both queries and reports are described in a structured manner defined by an XML Schema. The query is expressed in the gazetteer query language, which consists of Boolean combinations (AND, OR, and AND NOT) of seven types of queries. The following query types are combined for the purpose of our research:

- **name-query** *operator text*
  Returns all gazetteer entries having at least one name that matches *text* according to text-matching operator, *operator*. We employ the following operators for name-query type queries:
  - *Equals*. A gazetteer entry name matches the text if it equals text, ignoring insignificant differences in whitespace.
  - *Contains-phrase*. A gazetteer entry name matches text if it contains all of the words in the text in the same consecutive order. For example, entry name "Black Forest Drive" matches text "forest drive" under this operator, but entry names "Forest Lake Drive" and "Drive Forest" do not.
- **class-query** *thesaurus term*
  Returns all gazetteer entries belonging to class *term* or any subclass of *term* recursively (if the gazetteer supports subclasses or thesaurus relationships), where *term* is a term drawn from a thesaurus or simple vocabulary associated with the gazetteer. In our study, we consider as the thesaurus the "ADL Feature Type Thesaurus" and as the class term in which a place name may belong to one of the three: "administrative areas," "islands," or "historical sites."
- **relationship-query** *relation target identifier*
  Returns all gazetteer entries having a relationship *relation* to a target gazetteer entry identified by a *target identifier*. In our study, we make use of the relationship query whenever we need to narrow down the search to a placename within a specific country (that is, either Austria or Hungary). In that case, we apply the relationship "part of" to a target gazetteer entry identified by the corresponding country identifier code.

Based on these three query types, we ran the set of sample queries *four* times. The first time we used the "part of" relationship of the relationship query to narrow down the search to a placename within a specific country (that is, either Austria or Hungary) and the second time we did not use the relationship query at all and searched for the placename in the global scope only (that is, worldwide). In order to cope with short forms of composite names, for each case, we first tried the equality of the source term with a gazetteer term (that is, "equals" was defined as the name query operator) and then occurrence of the source term as part of the gazetteer term (that is, "contains-phrase" was defined as the name-query operator). The combined queries produced a reasonable match rate and are described below:

- *part of, equals*. A gazetteer query that AND-combines the class-query, the relationship-query, and the name-query attributed with the operator "equals."
- *part of, contains-phrase*. A gazetteer query that AND-combines the class-query, the relationship-query, and the name-query attributed with the operator "contains-phrase."

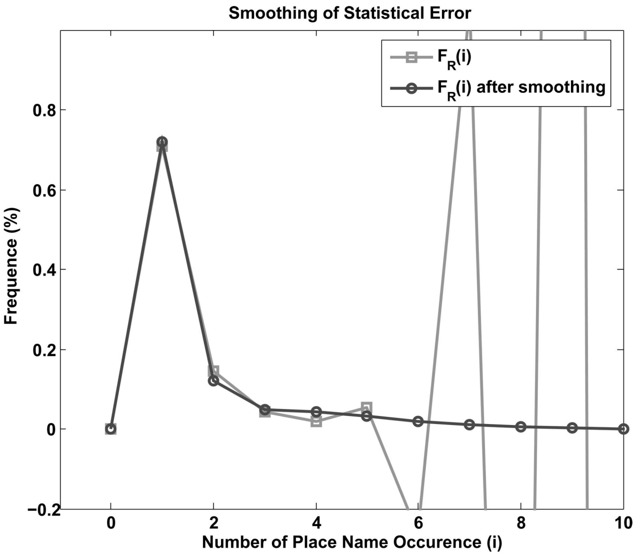

Fig. 2. Values of the vector $\vec{F}_R$ as they are computed before and after smoothing the statistical error in $\vec{F}_G$.

- *NO part of, equals*. A gazetteer query that AND-combines the class-query and the name-query attributed with the operator "equals."
- *NO part of, contains-phrase*. A gazetteer query that AND-combines the class-query and the name-query attributed with the operator "contains-phrase."

For each placename query, we register the gazetteer occurrence for this placename, in other words, the number of associations between places and placenames that we found in G for the given placename. At the end of the matching run, we compute all values of the $\vec{F}_G$ vector as approximated by the sampling. Then, we give arbitrary values to the constant probability P_{ASSOC} and compute the $\vec{F}_R$ vector. We adjust P_{ASSOC} until F_{0_R} falls to zero, following the assumption made in the methodology section that there are no placenames without places in the real world. Then, the solution for $\vec{F}_R$ provides an estimation of all of the placenames occurrences in the real-world structure.

4.2.1 Smoothing of the Sampling Data

The linear equation system we have to solve has a strongly dominant diagonal. Such systems are numerically stable, that is, they are insensitive to rounding and input data errors. Nevertheless, for the high values of placename occurrences, we experience some numerical instability due to statistical fluctuations in our limited sample. For example, for placename occurrence with a probability of appearing in our sample below, we can observe only one or zero. That is, the ratio of deviation from the expected value becomes very large. Therefore, for the higher values, the solution tends to oscillate between positive and impossible negative values. As shown in Fig. 2, this has little effect on the lower occurrence rates due to the form of the equation system and the absolutely small values of the observed higher occurrence rates. Further, only the lower occurrence rates contribute significantly to the parameters we are interested in, that is, semantic inconsistency, completeness of the gazetteer, recall, and precision.

It is reasonable to assume that the distribution of placename occurrences in the real-world structure also follows a statistical model. We therefore use a binomial distribution function to smooth this instability (statistical error) for the number of placename occurrences greater than four in order to investigate the effect of the instability on the results below. That is, we replace $\vec{F}_G$ for occurrences greater than four with a distribution with the same integral as $\vec{F}_G$ and fitting the frequency for occurrence 4. The values of $\vec{F}_R$ before and after the smoothing are depicted in Fig. 2. 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. Nevertheless, we attribute higher fidelity to the solutions from the smoothed values, so the rest of the experiments are based on them.

4.3 Experimental Results

Fig. 3 shows how the values of $\vec{F}_R$ in the four runs compare. Figs. 4, 5, 6, and 7 show the calculated values of $\vec{F}_R$ against the observed values of $\vec{F}_G$ for the four different runs.

Based on the values of the $\vec{F}_R$ vector, we compute the metrics of (2), (4), (5), (6), (7) defined in the previous section for the four runs. The results are given in Table 1.

Results indicate that

- The *Completeness* of ADL Gazetteer as estimated by the combined queries is approximately 65 percent for the target area Austria plus Hungary and the application characteristics.
- Knowledge of the identity of higher levels in a hierarchy of places does not seem to particularly influence our notion of completeness of the ADL Gazetteer (note that it is completeness with respect to all variant names, not with respect to all places).
- Knowledge of the identity of higher levels in a hierarchy of places results in a significant reduction in the semantic inconsistency, as one would expect since the placename multiplicity is lower. In particular, when the country of the place is known, the semantic inconsistency decreases from 8.13 percent to 3.83 percent in the case of the "*equals*" operator and from 12.8 percent to 6.85 percent in the case of the "*contains-phrase*" operator.
- Knowledge of the identity of higher levels in a hierarchy of places results in increasing of the precision of one-to-one mappings. Precision increases to 96.17 percent from 91.87 percent in the case of the "equals" operator and to 93.15 percent from 87.2 percent in the case of the "contains-phrase" operator.
- Recall of one-to-one mappings is 78.55 percent and 69.35 percent when the "part of" relationship is used and is 66.79 percent and 58.83 percent when the "part of" relationship is not used, for the cases of equals and contains-phrase, respectively. That is, by relaxing the matching criteria from, for instance, "equals" to "contains phrase," we produce more cases with duplicate names under the respective criterion. Since we require a simultaneously unique match, "relaxation" is actually a stronger constraint and causes a reduction of success rate and recall. Interesting enough, it also causes a reduction of

TABLE 1
Metrics of Completeness, Semantic Inconsistence, Success Rate, and Precision and Recall of One-to-One Mappings for the Four Runs

	PART OF, EQUALS	PART OF, CONTAINS-PHRASE	NO PART OF, EQUALS	NO PART OF, CONTAINS-PHRASE
$Comp_G$	0.6824	0.648	0.6522	0.6126
$G_{\substack{Semantic \\ Inconsistence}}$	0.0383	0.0685	0.0813	0.1280
$\text{SuccessRate}_{\substack{one-to-one \\ mapping}}$	0.5360	0.4494	0.4356	0.3604
$\text{Precision}_{\substack{one-to-one \\ mapping}}$	0.9617	0.9315	0.9187	0.8720
$\text{Recall}_{\substack{one-to-one \\ mapping}}$	0.7855	0.6935	0.6679	0.5883

precision due to higher placename frequencies, as explained by this theory.

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

4.4 Model Validation

To test the validity of our model, we exclude a number of places from the gazetteer data set that were identified throughout the mapping process and reapply the model to the residual data set to figure an *estimate* of the completeness, $Comp_G^{EST}$, of the gazetteer. Then, we compute the *actual* or *expected* completeness, $Comp_G^{EXP}$, of the gazetteer taking into account the original completeness, $Comp_G$, and the known reduction from excluding places. The difference between the estimate and the expected value, given in the form of *absolute error* and *percentage error*, is reported to provide experimental evidence that the model is sensitive to the gazetteer completeness as theoretically expected. Next, we describe the process of excluding places from the gazetteer.

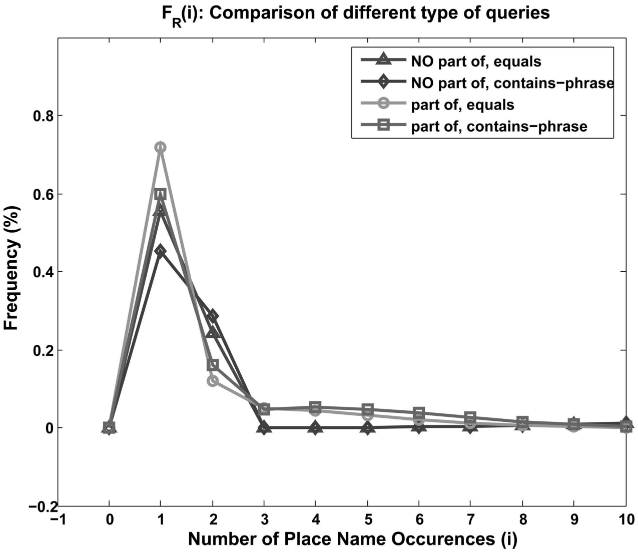

Fig. 3. Comparison of the different type of queries based on the values of the computed vector $\vec{F}_R$.

In the place removal process, we randomly choose a number of places to be removed from the original data set, compute the derived frequency distribution of the residual real placename occurrences, and define the standard deviation of it. *Simple random sampling without replacement* is used to determine the set of random places to be removed from the original gazetteer data set. This is a method of selecting n items out of the N such that every one of the n distinct samples has an equal chance of being drawn. In practice, a simple random sample is drawn item by item. The items in the population (that is, original gazetteer data set) are numbered from 1 to N. A series of random numbers between 1 and N is then drawn by means of a computer program that produces a table of random numbers. At any draw, the process used must give an equal chance to select any number in the *remaining* population. The items that bear these n numbers constitute the sample. The method is called "without replacement" because a number that has been drawn is removed from the population for all subsequent draws. For the distribution that results after randomly removing places from the original data set, the mean and the standard deviation are reported. The mean $\bar{x}$ and the standard deviation S of a frequency distribution are given by

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^{n} f_i}} \text{ and } \bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i},$$

where x_i stands for each data value in turn and f_i is the frequency with which data value x_i occurs. The mean and the standard deviation will show in Table 2 how the placename multiplicity and the distribution of placename multiplicity become smaller after removing places from the test set. Without loss of generality, in the next paragraph, we describe an example of the validation process for the case of *part of, equals* run. The same process is then followed for the other cases.

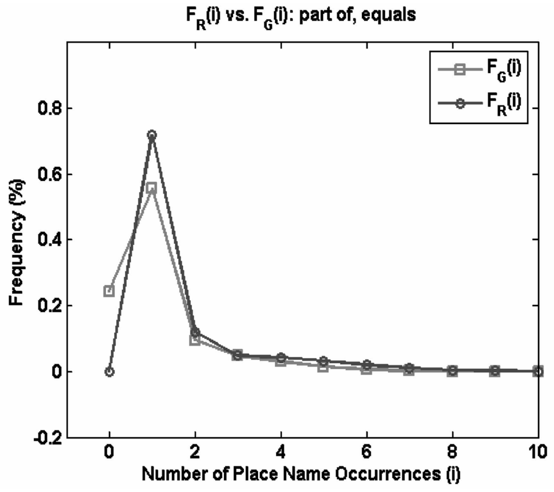

Fig. 4. $\vec{F}_R$ against $\vec{F}_G$: part of, equals.

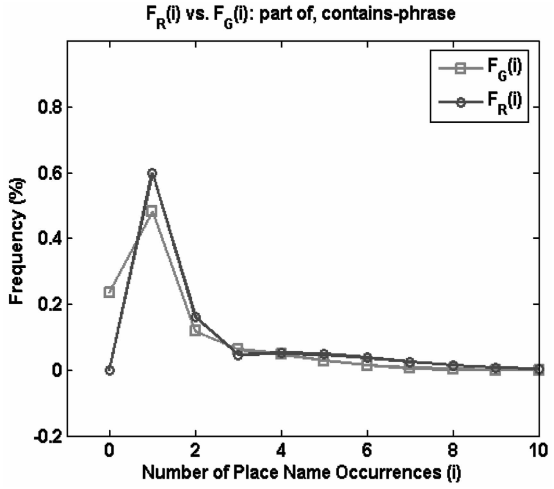

Fig. 5. $\vec{F}_R$ against $\vec{F}_G$: part of, contains-phrase.

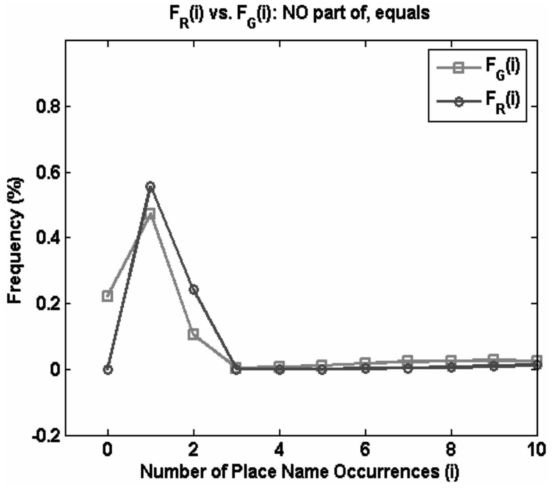

Fig. 6. $\vec{F}_R$ against $\vec{F}_G$: NO part of, equals.

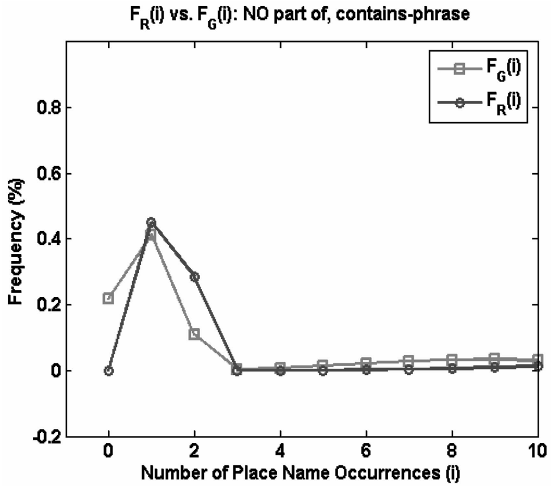

Fig. 7. $\vec{F}_R$ against $\vec{F}_G$: NO part of, contains-phrase.

In the case of *part of, equals* run, we found in the experimental results section (Section 4.3) that the completeness of the gazetteer is 0.6824 (that is, $Comp_G = 0.6824$). Also, from the mapping process, it is known that the sample data set of approximately 1,000 placenames has references to 1,327 places in the gazetteer (that is, $Card(PN_G) = 1,327$). Since $Comp_G = \frac{Card(PN_G)}{Card(PN_R)}$, it turns up that $Card(PN_R) = 1,944$. Now, consider that 10 percent of the gazetteer places are randomly removed (that is, 133 from the 1,327). Note that the removal process alters the frequency distribution of the real placename occurrences (for example, a placename that had references to three places may now have references to two places because one of the three places was excluded in the removal process). The *expected* completeness, $Comp_G^{EXP}$, is computed directly by $Comp_G = \frac{Card(PN_G)}{Card(PN_R)}$, where, now, $Card(PN_G)$ contains the places that were not excluded, whereas $Card(PN_R)$ remains the same. The *estimate* completeness $Comp_G^{EST}$ is computed by applying our model in the new frequency distribution. Table 2 shows the results of the validation process for the four different run methods when 10 percent, 20 percent, and 30 percent of the places are removed, respectively.

The validation results, as shown in Table 2, indicate that the error of our model for each case is below 4 percent, with an average percentage error over all cases of 2.04 percent. The errors reported are considered small and will succeed in validating the main assumptions made by our model. Furthermore, a slight deviation from our model is expected due to the random sampling process.

4.5 Discussion of Systematic Errors

First, we would like to stress that the precision of an error estimate below 10 percent of the computed error is hardly of any practical importance. For instance, the difference between an error of 4.5 percent to 5.0 percent means that the result is correct to 95.0 percent or 95.5 percent, that is, an effect of 0.005. Since the error itself is a statistical value of limited precision, such a variation disappears in practice behind statistical fluctuations.

Further, an error estimate should always be more on the pessimistic side. An error estimate serves to assess the reliability of results, for instance, if someone publishes statistics that are based on a placename matching process. If

TABLE 2
Validation of the Model for the Four Run Methods

RUN METHOD	$Comp_G$	% OF PLACES REMOVED	MEAN ($\overline{x}$)	STDEV (S)	$Comp_G^{EXP}$	$Comp_G^{EST}$	ABSOLUTE ERROR	PERCENTAGE ERROR (%)
PART OF, EQUALS	0.6824	10%	1.398	2.785	0.614	0.6325	0.0185	3.01
		20%	1.242	2.483	0.5456	0.5337	0.0119	2.18
		30%	1.087	2.331	0.4772	0.4809	0.0037	0.78
PART OF, CONTAINS-PHRASE	0.648	10%	1.751	3.463	0.5829	0.5599	0.023	3.95
		20%	1.556	3.077	0.5182	0.4987	0.0195	3.76
		30%	1.362	2.651	0.4534	0.4478	0.0056	1.24
NO PART OF, EQUALS	0.6522	10%	2.794	7.418	0.5869	0.5833	0.0036	0.61
		20%	2.485	6.590	0.5218	0.5214	0.0004	0.08
		30%	2.173	5.734	0.4564	0.4733	0.0169	3.7
NO PART OF, CONTAINS-PHRASE	0.6126	10%	4.141	11.826	0.5513	0.5414	0.0099	1.8
		20%	3.681	10.452	0.49	0.4952	0.0052	1.06
		30%	3.220	9.184	0.4288	0.4187	0.0101	2.36

we cannot precisely model a reality, we must make sure that the worst case is tolerable because the worst case can happen and not assume a possible better world. Therefore, an error estimate must better provide upper bounds in order to have practical value. This is good practice in data evaluation and numerical computation. We regard the assumption for our model to be realistic. That is, there are enough real cases in which the computed values are very close to reality. We show, in this paragraph, that some systematic effects may change to actual error to the better side. In this case, our model is slightly pessimistic, that is, our model is expected to be correct or at least on the "safe side."

Second, this paper aims at presenting the principles and not at a specific example since particular effects that may need to be taken into account will differ from application to application. For those readers not familiar with statistical models, however, it should be noted that the more different the effects that occur in a real-world system are, the better the stochastic model ignoring them all is at describing reality. Only extreme, dominant effects would really have a bearing, such as placename registration policies that address explicitly placename multiplicity. Under these premises, we now discuss the most relevant sources of systematic errors in our model.

4.5.1 Gazetteer Completion Strategies

Since the effect we are looking for has to do with the multiplicity of placename occurrence, we expect that only a significant deviation in the independence of P_{ASSOC} from the placename multiplicity would affect the results. The only reasonable strategy of a gazetteer curator we can imagine that would violate the assumption is the following: Let us assume that the gazetteer curators would systematically take the submission of a new placename as an occasion to find more places with that name in its domain. This would systematically increase the placename occurrence in the Gazetteer with respect to the real-world structure since places with rarer names would be less likely to be registered. Consequently, our model would overestimate the contribution of the higher placename occurrences and the semantic inconsistency. Under this consideration, our model would yield a more pessimistic inconsistency estimation.

Another strategy may be to register only one placename per place and to select a variant that is, if possible, unique. In this case, our model may yield too optimistic an inconsistency estimation. However, such a gazetteer should be regarded as not fit for the purpose of placename matching and can be easily recognized as such.

4.5.2 Chance to Hit the Wrong Place

Equation (4) about semantic inconsistency assumes that the chance to hit the wrong place in a gazetteer that occurs once in the gazetteer and n-times in the real world is statistical. If, however, our sample and the Gazetteer reveal a tendency to select the larger places, then one might expect that even the ambiguous cases contain more relevant hits than estimated. Again, our model would be more pessimistic with respect to the inconsistency.

4.5.3 Spelling Errors

Initially, our data contained systematic "spelling errors": Hungarian and German special characters, such as "ö," were mapped by some tools we used to make incorrect characters (Greek ones, actually). The Alexandria Gazetteer, on the other hand, uses nonstandard transcriptions to ANSI characters, for example, "ö" is transcribed to "o" instead of "oe." From 980 placenames, 255 contained such a spelling different from the representation in the Alexandria Gazetteer. Since these differences were due to a dozen rules, we could do the necessary transformations automatically. After carefully cleaning up such misspellings to the representation used in the Alexandria Gazetteer, we had no reason to assume a relevant percentage of nonexisting names in our samples. The case provided a nice test of this part of the theory: When trying to fit F_{0_R} to zero for the uncleaned sample, the equation system yielded "impossible" results, that is, diverging values greater than 1 and negative frequencies, as presented in Table 3. Obviously, the method is sensitive to data not conforming with the assumptions.

TABLE 3
Values Yielded by the Equation System for Cleaned and Uncleaned Data

NUMBER OF OCCURENCES	0	1	2	3	4	…
CLEANED	0.00	0.763	0.145	0.044	0.01748	…
UNCLEANED	0.00	15.41	-356.6	5514.9	-62268.1	...

Future work may try to extend the model so that even spelling errors can be estimated. This might be possible by reducing the degree of freedom in the equation system by introducing a statistical model of the possible shapes of the real-world distributions of placename occurrences, rather than having all frequencies of the real-world distribution besides F_{0_R} as unknowns. The idea is that a distorted signal must exhibit some similarity to the undistorted. This similarity could be further formalized and be exploited as a constraint. For our current results, we would expect that some residual spelling errors would slightly reduce the computed completeness of the gazetteer (in 1 percent ranges). This would again make our model slightly more pessimistic with respect to the inconsistency.

Summarizing, we expect that our results are reasonably close to reality and/or tend slightly to the pessimistic, the "safe" side. Sensitivity analysis on the above described effects and more elaborate models may further refine the presented method as indicated.

5 CONCLUSIONS

Digital information systematically contains references to locations or objects in geographic space. Information integration processes employ digital gazetteers to effectively 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 placenames is not isomorphic, the mapping process suffers from failures due to nonexistence, ambiguity, or semantic inconsistency. In this work, we analyzed the cases of success and the cases for failure in the mapping process between a geographic name given by the text in an uncontrolled data field of a source and the controlled term of the identical geographic name described in a digital gazetteer. We presented a statistical model that permits estimating the degree of completeness of a gazetteer and the expected quality of a set of placename mappings to a gazetteer, based on observation of mappings themselves without the need for further data. We estimated the completeness of the ADL gazetteer to be about 65 percent for the area of Austria and Hungary that our test data covers. Our experiments show that knowledge of the identity of higher levels in a hierarchy of places improves the precision of one-to-one mappings since Precision increases to 96.17 percent from 91.87 percent in the case of the "equals" operator and to 93.15 percent from 87.2 percent in the case of the "contains-phrase" operator. Furthermore, it results in a significant reduction of the semantic inconsistency, as one would expect, since the placename multiplicity is lower. In particular, when the country of the place is known, the semantic inconsistency decreases from 8.13 percent to 3.83 percent in the case of the "equals" operator and from 12.8 percent to 6.85 percent in the case of the "contains-phrase" operator.

As far as we know, our work is the first that provides estimations of the quality of mappings to a digital gazetteer and its completeness with respect to the application. We expect our method to be found useful in various upcoming studies. We regard that the presented method can be useful for

- determining the error propagation introduced by mismatch into statistical analysis based on the matched data such as "how many objects/people/events of kind X were 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, and
- providing decision support for gazetteer use and gazetteer development issues.

The methodology we describe is based on assumptions that allow the formulation of a mathematical deconvolution problem. We have discussed ways in which the model could be modified if the assumptions are not sufficiently fulfilled or if a higher precision of results is sought. This approach is the main contribution of this paper. The methodology we describe is general and can probably be easily adopted to matching terms against other types of well-defined concepts.

ACKNOWLEDGMENTS

This work was supported in part by the Ubi-Erat-Lupa project under the Culture 2000 program of the European Union.

REFERENCES

[1] D. Calvanese, S. Castano, F. Guerra, D. Lembo, M. Melchiori, G. Terracina, D. Ursino, and M. Vincini, "Towards a Comprehensive Methodological Framework for Semantic Integration of Heterogeneous Data Sources," *Proc. Eighth Int'l Workshop Knowledge Representation Meets Databases (KRDB '01),* 2001.

[2] *Intelligent Integration of Information,* G. Wiederhold ed. Kluwer Academic, 1996.

[3] S. Bergamaschi, S. Castano, S. De Capitani di Vimercati, S. Montanari, and M. Vincini, "An Intelligent Approach to Information Integration," *Formal Ontology in Information Systems,* N. Guarino ed., IOS Press, 1998.

[4] J. Artz, "A Crash Course in Metaphysics for the Database Designer," *J. Database Management,* vol. 8, no. 4, 1997.

[5] E. Remolina, J. Fernández, A. Kuipers, and J. González, "Formalizing Regions in the Spatial Semantic Hierarchy: An AH-Graphs Implementation Approach," *Proc. Conf. Spatial Information Theory (COSIT '99),* 1999.

[6] L.L. Hill and M. Goodchild, "Digital Gazetteer Information Exchange (DGIE)," final report of workshop, http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/DGIE%20final%20report.htm, 2000.

[7] L.L. Hill and Q. Zheng, "Indirect Geospatial Referencing through Placenames in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers," *Proc. 62nd Ann. Meeting Am. Soc. for Information Science,* 1999.

[8] L.L. Hill, "Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints," *Proc. Fourth European Conf. Digital Libraries (ECDL '00),* 2000.

[9] Alexandria Digital Library Project, http://www.alexandria.ucsb.edu, 2005.

[10] D. Bitton and D.J. DeWitt, "Duplicate Record Elimination in Large Data Files," *ACM Trans. Database Systems (TODS),* vol. 8, no. 2, 1983.

[11] V.S. Verykios, A.K. Elmagarmid, and E.N. Houstis, "Automating the Approximate Record-Matching Process," *J. Information Sciences,* vol. 126, nos. 1-4, 2000.

[12] M.A. Hernández and S.J. Stolfo, "Real-World Data Is Dirty: Data Cleansing and the Merge/Purge Problem," *Data Mining and Knowledge Discovery J.,* vol. 2, no. 1, 1998.

[13] T.W. Yan and H. Garcia-Molina, "Duplicate Removal in Information Dissemination," *Proc. Very Large Databases (VLDB '95),* 1995.

[14] M.L. Lee, H. Lu, T.W. Ling, and Y.T. Ko, "Cleansing Data for Mining and Warehousing," *Proc. 10th Int'l Conf. Database and Expert Systems Applications (DEXA '99),* 1999.

[15] C.A. Knoblock, S. Minton, J.L. Ambite, N. Ashish, P.J. Modi, I. Muslea, A.G. Philpot, and S. Tejada, "Modeling Web Sources for Information Integration," *Proc. 15th Nat'l Conf. Artificial Intelligence (AAAI '98),* 1998.

[16] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Data Integration in Data Warehousing," *Int'l J. Cooperative Information Systems,* vol. 10, no. 3, pp. 237-271, 2001.

[17] W.W. Cohen, "Knowledge Integration for Structured Information Sources Containing Text," *Proc. SIGIR-97 Workshop Networked Information Retrieval,* 1997.

[18] D.A. Smith and G. Crane, "Disambiguating Geographic Names in a Historical Digital Library," *Proc. Fifth European Conf. Research and Advanced Technology for Digital Libraries (ECDL '01),* 2001.

[19] OpenGIS Consortium, *The OpenGIS Abstract Specification Topic 5: Features,* Version 4, OpenGIS Project Document Number 99-105r2, C. Kottman, ed., http://portal.opengeospatial.org/files/?artifact_id=890, 24 Mar. 1999.

[20] A. Mikheev, M. Moens, and C. Grover, "Named Entity Recognition without Gazetteers," *Proc. Ninth Conf. European Chapter of the Assoc. for Computational Linguistics (EACL '99),* 1999.

[21] R. Bunescu and M. Pasca, "Using Encyclopedic Knowledge for Named Entity Disambiguation," *Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics (EACL '06),* 2006.

[22] R.K. Srihari, W. Li, T. Cornell, and C. Niu, "InfoXtract: A Customizable Intermediate Level Information Extraction Engine," *J. Natural Language Eng.,* vol. 12, no. 4, pp. 1-37, 2006.

[23] M. Bilenko and R.J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," *Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* 2003.

[24] T.W. Yan and H. Garcia-Molina, "Duplicate Removal in Information System Dissemination," *Proc. 21st Int'l Conf. Very Large Data Bases (VLDB '95),* 1995.

Martin Doerr studied mathematics and physics from 1972 to 1978 and received the PhD degree in experimental nuclear physics from the University of Karlsruhe, Germany. His industrial experience was the successful product development of a multiprocessor operating system at Bruker, Karlsruhe, Germany, from 1984 to 1990. Since 1990, he has been a principle researcher at FORTH. He has been leading or participating in a series of national and international projects for knowledge management, terminology management, cultural information systems, and information integration systems. He is leading the working group of the International Committee for Documentation of the International Council of Museums (ICOM/CIDOC), which has developed ISO21127:2006, together with the respective International Organization for Standardization (ISO) committees, a standard core ontology for the semantic interoperability of cultural heritage information. His research interests include ontology engineering, semantic interoperability, and information integration. He has published more than 30 refereed papers and six book chapters. He is a member of the editorial board of the journal *Applied Ontology* and the *ACM Journal on Computers and Cultural Heritage*.

Manos Papagelis received the BSc and MSc degrees in computer science from the University of Crete, Greece, in 2002 and 2005, respectively. Since 2005, he has been a PhD candidate in computer science at the University of Toronto, Canada. From 2000 to 2002, he was a member of the Atlantis Group Research Unit at the University of Crete and, from 2002 to 2005, a research fellow at the Institute of Computer Science, FORTH. He has published in international refereed conferences and journals and he is the software architect of the "Confious" (www.confious.com) conference management system that has successfully served a number of academic and scientific conferences. His research interests include Web search, collaborative filtering, text analysis and mining, and network data management.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.