A data model for efficient address data representation – Lessons learnt from the Intiendo address matching tool

Abdullah Al Rahed[1], Serena Coetzee[2], Magnus Rademeyer[3]

[1]SDSL, Dhaka, Bangladesh, rahed@afrigis.co.za
[2]Dept Computer Science, University of Pretoria, Pretoria, South Africa, scoetzee@cs.up.ac.za
[3]AfriGIS, Pretoria, South Africa, magnus@afrigis.co.za

Abstract

Geocoding refers to the process of assigning geographic identifiers and/or geographic coordinates to the description of a feature location, i.e. the words, codes or terms that describe a feature's location. Address matching is the specific case of geocoding where the description of a feature location comprises an address. Address matching is complicated by an incomplete or inaccurate incoming address or one that contains a misleading geographic identifier in its location type hierarchy. The Intiendo address matching tool is based on a data structure that is similar to the hierarchy of location types described in ISO 19112, but we have made some novel extensions to enable a spatial adjacency search. Intiendo does not rely purely on the alphanumeric match in the location type hierarchy but incorporates spatial proximity into the address matching. We also describe how the Intiendo address matching process can be configured and fine tuned, for example, by assigning weights to the location types in the hierarchy, and by specifying parameters for the spatial adjacency match. In this paper we present the hierarchical data structures of the Intiendo address matching tool and show how they are an extended implementation of the ISO 19112 general model. We show the similarities between the Intiendo and ISO 19112 models, and present the extensions that were implemented in Intiendo. By way of examples, we show that our extended model allows more efficient and accurate address matching incorporating spatial adjacency and hierarchical fine tuning.

1. Introduction

Geocoding refers to the process of assigning geographic identifiers and/or geographic coordinates to the description of a feature location, i.e. the words, codes or terms that describe a feature's location. Address matching is the specific case of geocoding where the description of a feature location comprises an address. A geocoding service receives as input the description of the feature location, such as an address, and searches for a matching address in a reference dataset.

Addresses are often structured into a spatial hierarchy that describes a location with increasing accuracy. In the address '14 Richmond Road, Mowbray, Cape Town, South Africa' the spatial accuracy increases from country (South Africa) to city (Cape Town) to suburb (Mowbray) to street

(Richmond Road) to street number (14). The international standard, ISO 19112 – *Spatial referencing by geographic identifiers*, provides a general model for spatial referencing using geographic identifiers and defines the components of a spatial reference system. In the ISO 19112 model, a spatial reference system using geographic identifiers comprises a related set of one or more location types, together with their corresponding geographic identifiers, and the location types may be related to each other through aggregation or disaggregation, possibly forming a hierarchy. This general model is applicable to an address structured into a spatial hierarchy, as in our address above: country, city, suburb, street and street number are the location types, each with their own set of geographic identifiers, e.g. city names of South Africa are the geographic identifiers of the 'city' location type, and 'Cape Town' is a geographic identifier of that location type describing a specific location.

The process of address matching is complicated by an input address that is incomplete or inaccurate, or one that contains a misleading geographic identifier in its location type hierarchy. The cause of such an input address is often due to the ambiguities originating from uncertainties regarding suburb boundaries. While a single set of official place name boundaries for a country can reduce such ambiguities, Coetzee and Cooper (2007) point out that one will never get rid of these ambiguities because suburb or place name boundaries are not the type of boundary to be physically fenced off and hence obvious for all to see. There is also always the human ego factor that sees a person, living near the boundary of a more prestigious suburb, use the name of that suburb in their address. An example of an input address with a misleading geographic identifier in its location type hierarchy is '101 Rubida Street, Murrayfield'. A geocoding algorithm that employs a top-down alphanumeric matching approach based on the location type hierarchy will incorrectly match this address to '110 Rubida Street, Murrayfield', and not to the more accurate '101 Rubida Street, Die Wilgers' on the opposite side of the road. Refer to Figure 1.

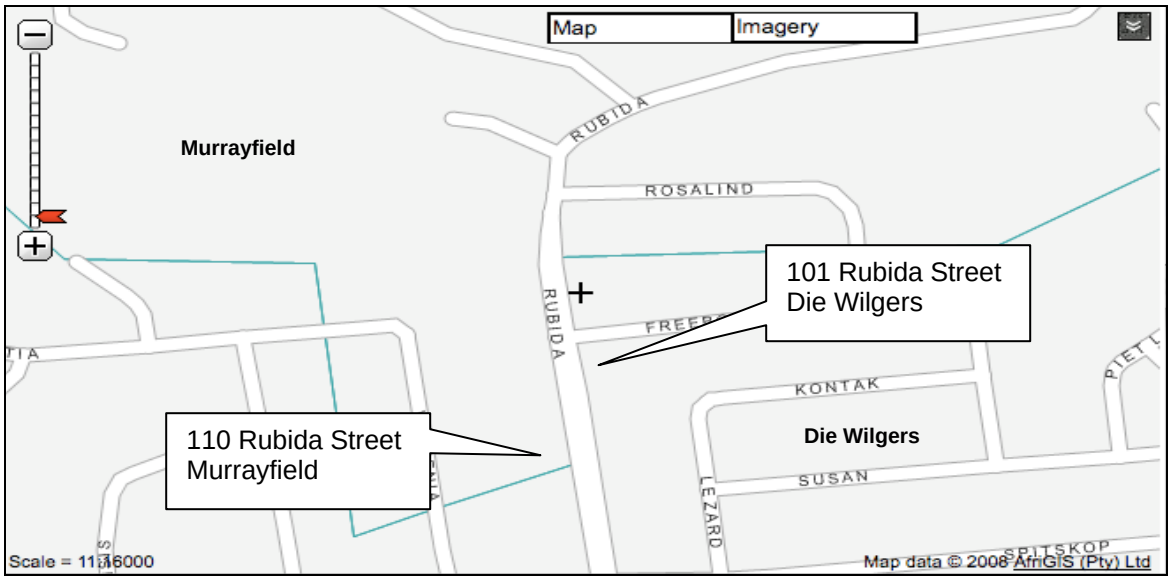

Figure 1: Rubida Street as the boundary between 'Murrayfield' and 'Die Wilgers

Relaxing the requirement to match the suburb accurately would not only add ‘101 Rubida Street, Die Wilgers’, to a list of potential matches, but also ‘101 Rubida Street, Rondebosch’ and ‘101 Rubida Street, Wilgenhof’, addresses in other parts of the country. The question is which one of these potential ‘101 Rubida Street’-addresses is the correct one? While ‘Die Wilgers’ and ‘Wilgenhof’ are closer in terms of string matching, they are spatially much further apart than ‘Murrayfield’ and ‘Die Wilgers’. In their survey on the field of geocoding, Goldberg et al. (2007) list attribute relaxation as one of the common causes of error in the matching stage of the geocoding process. Davis and Fonseca (2007) propose a so-called geocoding certainty indicator (GCI) that takes into consideration the spatial transformations that an address record goes through during the matching, and the approximations used to match the input address with an existing address in the reference dataset. This indicator considers alphanumeric proximity of suburbs (based on string matching) but not spatial proximity.

The Intiendo address matching tool is based on a data structure that is similar to the related set of location types described in ISO 19112, but we have made some novel extensions to enable a spatial adjacency search. Intiendo does not rely purely on the alphanumeric or string match against the set of the location types but incorporates spatial proximity into the address matching process so that the above address would be matched correctly to ‘101 Rubida Street, Murrayfield’. We also describe how the Intiendo address matching process can be configured and fine tuned, for example, by assigning weights to the location types in the hierarchy, and by specifying parameters for the spatial adjacency match.

In this paper we present the hierarchical data structures of the Intiendo address matching tool and show how they are a specific implementation of the ISO 19112 general model. We show the similarities between the Intiendo and ISO 19112 data models, and present the extensions that were implemented in Intiendo for address matching. By way of examples, we show that our extended model allows more efficient and accurate address matching incorporating spatial adjacency and hierarchical fine tuning.

2. The Intiendo data model

Intiendo (Spanish for ‘I understand’) is a software toolset that can be used to structure and geocode addresses, i.e. to understand and interpret addresses. Intiendo is based on the principle that an address has a hierarchical pattern, i.e. the street number belongs to a street, the street belongs to a suburb, town, province, etc. An Intiendo hierarchy consists of a number of levels, forming a hierarchy. Before address matching can take place, an address reference dataset is converted into a so-called Intiendo hierarchy database, which is structured according to a specific Intiendo hierarchy and contains the data (items) for each level of the hierarchy. Incoming addresses, either in free format or in a number of address lines, are parsed, structured and matched against the Intiendo hierarchy database. Refer to Figure 2 for an Intiendo hierarchy for addresses of the Street Address type described in the South African address standard (SANS 1883), and to Figure 3 for an excerpt of data from an Intiendo

hierarchy database based on the Intiendo hierarchy in Figure 2. In Intiendo the data for each level is referred to as individual items.

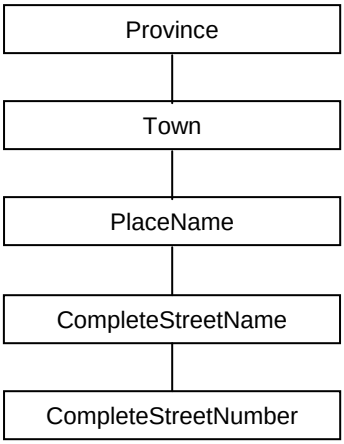

Figure 2. Intiendo hierarchy based on the SANS 1883 *Street address* type

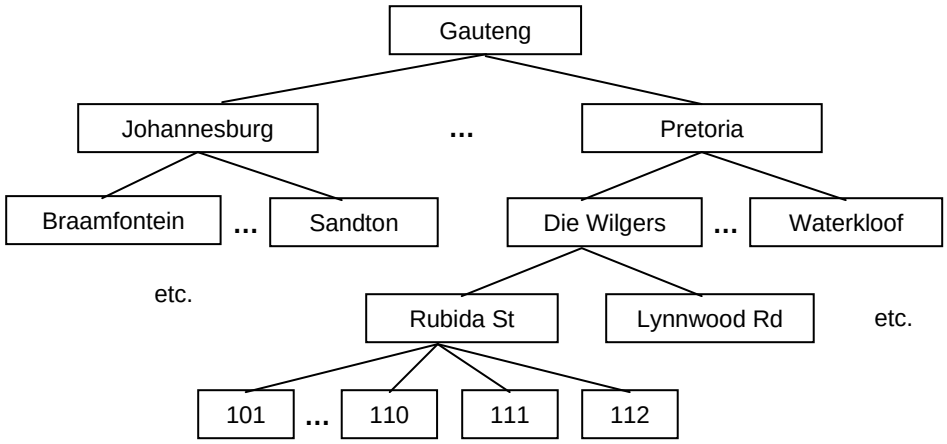

Figure 3. Sample data in an Intiendo hierarchy database based on the Intiendo hierarchy in Figure 1

Each item in an Intiendo hierarchy database has a number of default attributes, i.e. the item identifier (a unique identifier assigned by the Intiendo hierarchy builder), item code (a unique key taken from the reference dataset), item name, alternate names (other names by which the item is known) accuracy (metadata to describe the coordinate), and a longitude and latitude (geo-referenced location or approximation of the item). Figure 4 shows the Intiendo data model. It is also possible to add user-defined attributes as illustrated by the *PostalSuburb* and *StreetCode* attributes that are displayed in Figure 5.

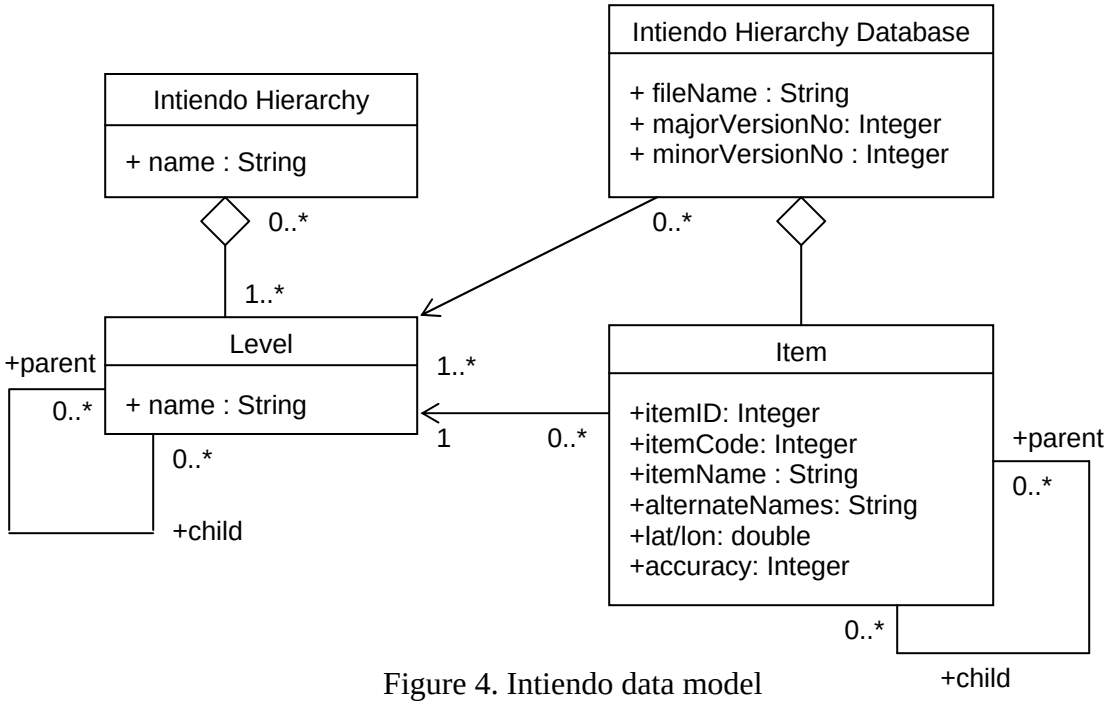

Figure 4. Intiendo data model

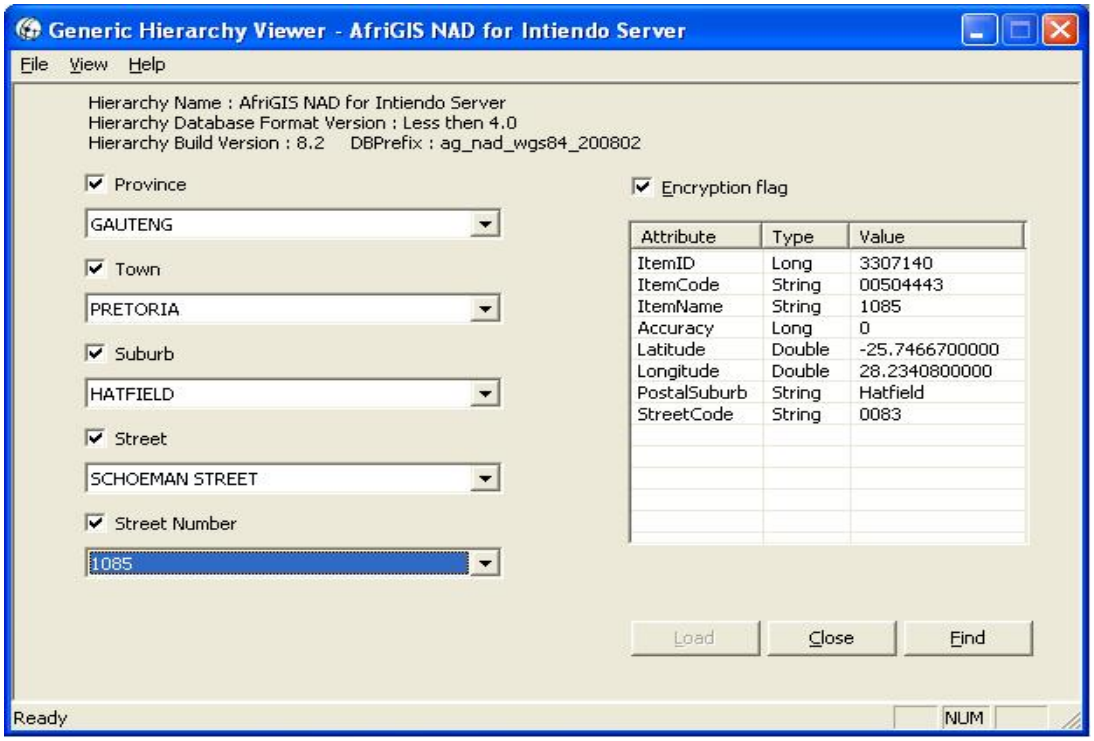

Figure 5. Sample Intiendo hierarchy database, presented in the Intiendo Hierarchy Viewer

Before the actual Intiendo address matching process starts, input addresses are parsed and structured according to the Intiendo hierarchy against which matching will take place, i.e. incoming address elements are detected and assigned to a level. This structured address data is then matched against an Intiendo hierarchy database (HDB) as illustrated in Figure 6. The first step of the address matching process is to do an *OptimizedMatch* of the structured input addresses. Depending on the outcome of this match, we either have a geocoded output address, or a further *SpatialAdjacencyMatch* is done in an attempt to find a more suitable match. The *OptimizedMatch* and *SpatialAdjacencyMatch* procedures are illustrated in Figures 7 and 8 respectively.

The *OptimizedMatch* implements the edit distance algorithm for alphanumeric string matching and the numeric distance matching algorithm for numeric input. An incoming address element is compared to the item names on each level of the Intiendo hierarchy database, and a corresponding matching percentage (MP) is calculated. The item with the highest MP is selected as the best possible address match. The *OptimizedMatch* procedure can be fine tuned by setting parameters such as specifying an anchor level, specifying the weight of a level, and specifying the matching percentage threshold of a level. If an anchor level is specified, the *OptimizedMatch* will first try to match incoming address elements on this level with items of the corresponding level in the Intiendo hierarchy database (HDB). Only the parents and children of those items in the HDB for which the MP is above the MP threshold of the anchor level are included in the subsequent search refinement.

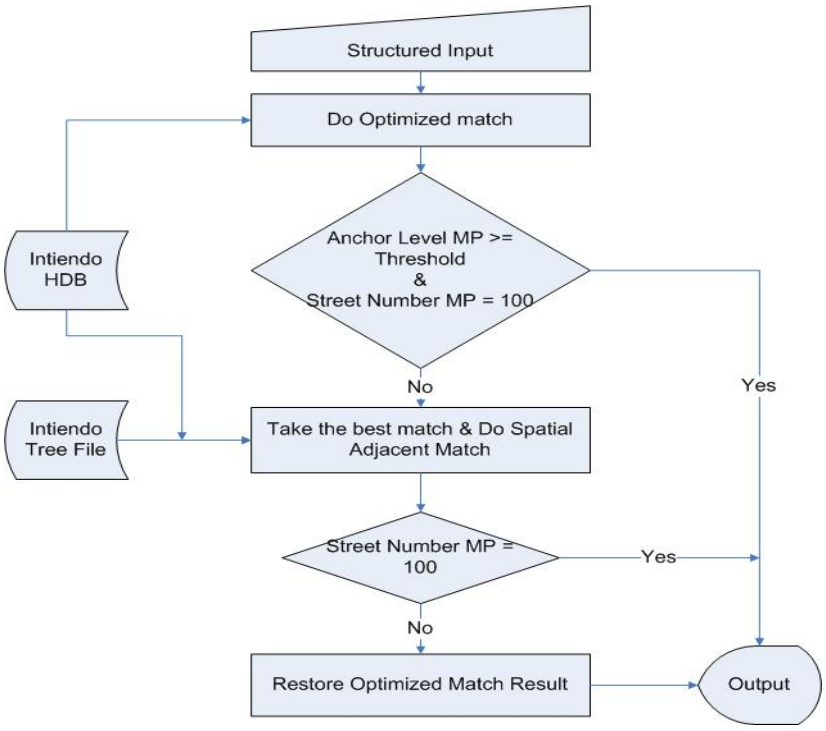

Figure 6. Intiendo address matching process

Giving preference to a single level of the Intiendo hierarchy does not always produce the desired result, and as an alternative approach one can prioritize the levels by assigning a weight to individual levels in the Intiendo hierarchy. In such a case, the *OptimizedMatch* calculates an average MP (AMP) based on the individual MPs on each level and their associated weights, i.e. all levels of the hierarchy are considered in the AMP and not only a single anchor level. This approach works well if it is known that a specific level of the incoming address is always empty, since assigning a 0 (zero) weight to this level will effectively remove its effect on the AMP. Goldberg *et al*. (2007) refer to this approach as attribute relaxation. The MP and AMP that are calculated during the *OptimizedMatch* consider alphanumeric proximity only and are therefore comparable to the CGI that Davis and Fonseca (2007) propose.

These different approaches to address matching are not universally applicable to any incoming address dataset, but largely depend on the type and quality of incoming address data. Sometimes it is necessary to split the incoming address data into different parts, and apply different approaches to different parts of the data. More often than not, geocoding is an iterative process of trying different approaches to different kinds of incoming addresses until the best possible overall match can be determined.

As an example, an incoming address of 1085 Schoeman Street, Hatfield, Randburg, Gauteng includes an incorrect item on the Town level, i.e. Randburg should be replaced with Pretoria. However, this incoming address will be matched accurately to 1085 Schoeman Street, Hatfield, Pretoria, Gauteng if setting their weights prioritizes the CompleteStreetNumber, CompleteStreetName and Suburb levels.

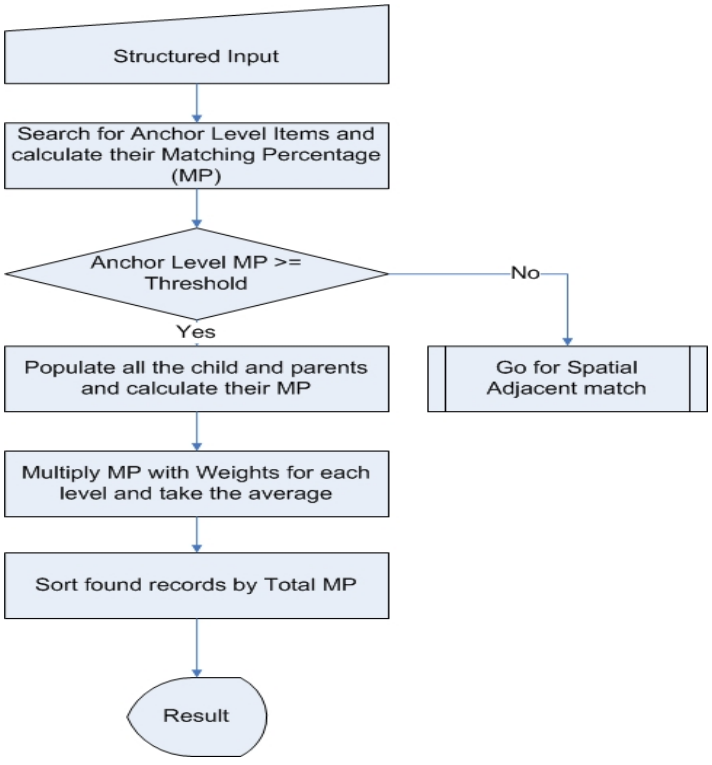

Figure 7. *OptimizedMatch*

OptimizedMatch successfully matches incorrect or incomplete incoming addresses by making use of the anchor level or weights per level. However, for misleading information as part of an incoming address, as illustrated in the example in the Introduction, a different kind of matching is required. As explained earlier, it is common to use a neighboring or adjacent place name in an address that is close to the boundaries of the two places. For this kind of problem Intiendo employs the novel *SpatialAdjacencyMatch*. As another example, *OptimizedMatch* will match an incoming address of *1001 Schoeman Street, Hatfield* to *1019 Schoeman Street, Hatfield* if the CompleteStreetNumber, CompleteStreetName and Suburb levels are prioritized because this street number is the best alphanumeric or string match within the boundary of the HATFIELD place name. However, *1001 Schoeman Street* can actually be found in the adjacent place of ARCADIA. Relaxing the attributes by prioritizing the CompleteStreetNumber and CompleteStreetName levels will also produce a wrong match since other places, far away from HATFIELD, could also have a *1001 Schoeman Street*. To overcome this problem those places that are within a specific radius from HATFIELD, having the specified CompleteStreetName and CompleteStreetNumber should be considered as potential matches. To ensure that the shape of a place name boundary does not preclude adjacent places, a kd-tree (*k* dimensional tree) space partitioning data structure is used for the spatial data representation.

The indexing structure of a kd-tree provides access to a set of data objects in the order of decreasing spatial proximity, thus reducing the number of redundant objects that have to be fetched, as well as the number of objects that have to be examined. Since it is expensive to perform spatial operations such as intersection and containment for the exact location and extent of a spatial feature, some initial approximations and filtering are done. The container approach is followed: the minimum-bounding rectangle (box) — the smallest rectangle (box) that encloses the object – is used to represent an object, and only when the test on the container succeeds is the real object examined.

SpatialAdjacencyMatch considers only those points within a calculated bounding box. In the above example, *SpatialAdjacencyMatch* calculates the bounding box from the center of the place in the best match provided by the *OptimizedMatch*, i.e. *1019 Schoeman Street, HATFIELD*. The kd-tree search returns the street numbers closest to the *1019 Schoeman Street, HATFIELD* address. Since *1001 Schoeman Street* is in the adjacent place, the kd-tree search results include this address. Thus, Intiendo finds *1001 Schoeman Street, Arcadia* as the best match for the incoming address of *1001 Schoeman Street, Hatfield*, effectively ignoring the place name level, even though it might be prioritized for the *OptimizedMatch*, but also eliminating spatially distant yet alphanumerically close matches. This correct match would not have been possible without taking spatial adjacency into consideration, and the use of the kd-tree for indexing ensures that the performance and efficiency of this spatial match are acceptable. Place name boundaries vary in size and shape and therefore Intiendo allows the user to specify the radius that is used to determine the bounding box from the center of the place. Currently the user has to set the radius per address matching run. In future the radius could be determined on the fly from the size of suburbs in the vicinity.

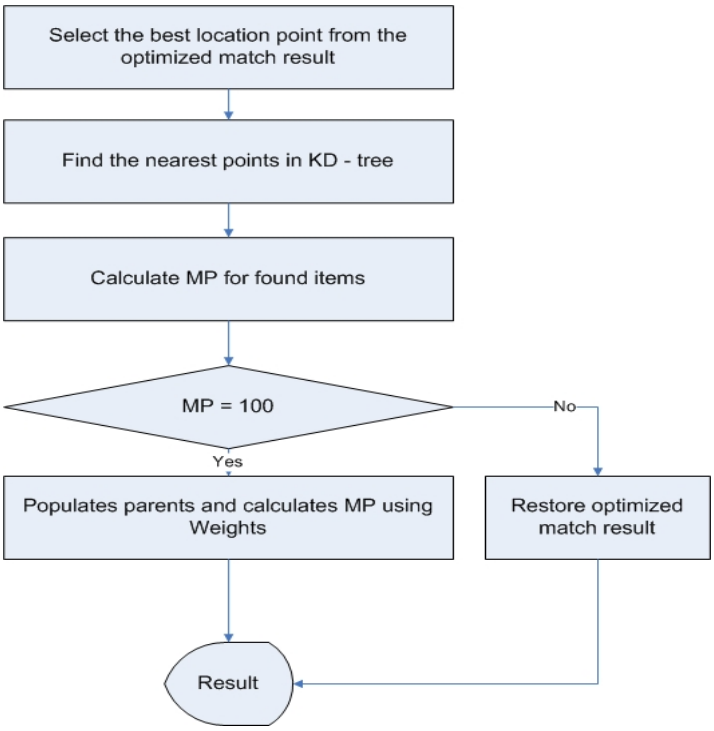

Figure 8. *SpatialAdjacencyMatch*

On a higher level, Intiendo also supports multi-hierarchy search, i.e. the user specifies a prioritized list of Intiendo hierarchy databases that should be used during address matching. If the matching percentage for an address is below a certain threshold for the first Intiendo hierarchy database and no spatial match is found, matching proceeds against the other Intiendo hierarchy databases in the list. The Intiendo address matching toolset has been successfully used to geocode large volumes (4 million address records or more per dataset) of diverse address datasets for a number of clients in both the private and the public sector.

3. The general data model for spatial referencing using geographic identifiers

Geographic information contains spatial references that relate the features and information represented in the data or text to positions in geographic space. Spatial references fall into two categories:

a) those using coordinates, and

b) those using geographic identifiers.

ISO 19111 - Geographic information - Spatial referencing by coordinates deals with the former while *ISO 19112 – Geographic information – Spatial referencing by geographic identifiers* deals with the latter, sometimes referred to as "indirect" spatial referencing. ISO 19112 defines the conceptual schema for spatial references based on geographic identifiers and establishes a general model for

spatial referencing using geographic identifiers; it defines the components of a spatial reference system and defines the essential components of a gazetteer.

According to ISO 19112, a spatial reference system using geographic identifiers comprises a related set of one or more location types, together with their corresponding geographic identifiers. These location types may be related to each other through aggregation or disaggregation, possibly forming a hierarchy. Refer to Figure 9. Examples of location types are a municipality, a town, a locality, or a street. The geographic identifiers of the *Street* location type, for example, are the street names, and an example of a location instance of the *Street* location type with territory of use in South Africa is 'Table Bay Boulevard'. A gazetteer is a directory of geographic identifiers describing location instances.

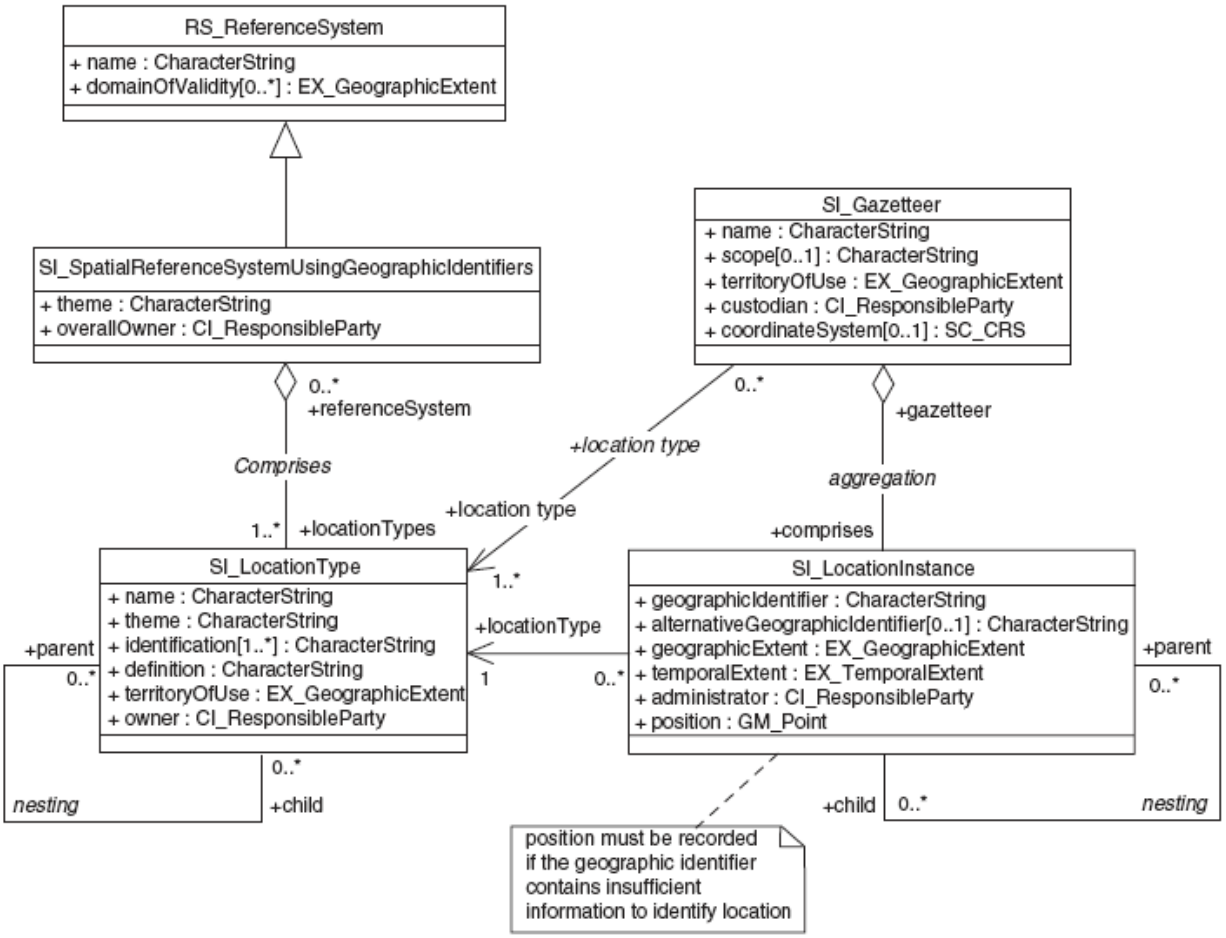

Figure 9. UML model for spatial reference system using geographic identifiers (ISO 19112, 2003)

4. Data model comparison

An addressing system is a specialization of a spatial referencing system by geographic identifiers as described in ISO 19112. According to ISO 19112, a *spatial reference* is a description of position in the real world (such as an address) and a *spatial reference system* is a system for identifying position in the real world (such as an addressing system). A *geographic identifier* is a spatial reference in the form of a label or code (such as a place name or a street name or an address) that identifies a location. A *spatial*

reference system using geographic identifiers is a system for describing positions in the real world with labels or codes and comprises a related set of one or more *location types* that may be related to each other through aggregation or disaggregation, possibly forming a hierarchy. A *gazetteer* is a directory of *instances of location types*.

An *address* is a spatial reference in the form of a hierarchically related group of geographic identifiers. An *addressing system* is a spatial reference system using addresses for describing position in the real world. It comprises a related set of one or more *location types* that usually form a hierarchy. An *address* is an instance of a valid group of hierarchically related location types, as allowed by the rules of the addressing system. An *address dataset* is a directory of *addresses*. Table 1 summarizes the comparison of these concepts.

Table 1. Concepts in ISO 19112, compared to the special case of an address

ISO 19112	**Special case of an address**
location identifiable geographic place	(same)
spatial reference description of position in the real world	(same)
geographic identifier a spatial reference in the form of a label or code	address a spatial reference in the form of a hierarchical group of geographic identifiers
spatial reference system using geographic identifiers a system for describing positions in the real world with labels or codes, comprising a related set of one or more location types *that may be related to each other through aggregation or disaggregation, possibly forming a hierarchy*	addressing system a system for describing position in the real world with addresses, comprising a related set of one or more location types *that usually form a hierarchy*
gazetteer a directory of instances of location types	address dataset (for reference purposes) a directory of instances of hierarchical groups of location types

In Table 2 the ISO 19112 column on the left provides the terminology for gazetteers in general, while the centre column relates this to the specific case of an address, and the right hand column (Intiendo) shows the terminology that is used in an implementation of this specific case. The purpose of this table is to assist in comparing and relating the ISO 19112 and Intiendo data models.

The Intiendo hierarchy, which represents a spatial reference system, is well-defined and comprises a set of location types with a common theme, a major requirement for conformance with a spatial referencing system using geographic identifiers, according to ISO 19112. Table 3 is a description, based on ISO 19112, of the location types used in an Intiendo hierarchy for addresses of South Africa. Intiendo hierarchies are also in conformance with ISO 19112 because the construction and attribute

data are well-defined and well-known. When an Intiendo hierarchy database is built, the Intiendo hierarchy builder ensures that all the instances of the location items are recorded in the gazetteer and the attribute data for each item is recorded correctly.

Table 2. Concepts in ISO 19112, compared to the special case of an address

ISO 19112	Special case of an address	Intiendo
Geographic identifier	An address consists of a group of hierarchically related geographic identifiers	An address consists of a group of hierarchically related items
Location type	Location type such as street, suburb, town, province, etc.	Hierarchy level
Spatial reference system using geographic identifiers	Addressing system	Intiendo hierarchy
Gazetteer	Address dataset	Intiendo hierarchy database

Table 3. Description of location types in Intiendo hierarchy database.

Name	Theme	Identifier	Definition	Territory of use	Owner	Parent	Child
province	administrative boundary	name	provincial authority	South Africa	Municipal Demarcation Board	none	town
town	colloquially known	name	colloquially known city or town	South Africa	AfriGIS	province	suburb
suburb	recorded name at a Surveyor General's office or used colloquially	name	suburb or place name	South Africa	Local authorities	town	street
street	access	name	road providing access to the service delivery point	South Africa	AfriGIS	suburb	street number
street number	service delivery point	unique street reference number	identifies the specific service delivery point	South Africa	AfriGIS	street	none

The hierarchical tree structure of an Intiendo hierarchy database ensures that a geographic identifier (item) is unique within a wider geographic domain. Each item in an Intiendo hierarchy database is assigned a unique identifier, the itemID, and each item also has coordinates associated with it, similar to the position (*GM_Point*) attribute of the *SI_LocationInstance* described in ISO 19112. An Intiendo hierarchy conforms not only to a spatial reference system using geographic identifiers as described in ISO 19112, but it also includes referencing by coordinates as described in ISO 19111. Similar to the

ISO 19112 data model, an Intiendo hierarchy database includes elements of both ISO 19112 and ISO 19111: every item in the Intiendo hierarchy database is not only represented by a group of geographic identifiers such as “1001 Schoeman Street, Arcadia, Pretoria, Gauteng”, but also by coordinates such as “-25.74694, 28.23019”. Without these coordinates, the spatial adjacency match would not be possible. The implementation of a kd-tree for each location type in Intiendo is an extension of the data model described in ISO 19112 and improves the quality and performance of search results, especially for the spatial adjacency search. The approximations and filtering described above ensure that the spatial match is performed within acceptable response times. The ability to set an anchor level and to prioritize Intiendo hierarchy levels during an address match allows for flexibility to fine tune the address matching process for different kinds of input address data.

5. Conclusion

In this paper we presented the hierarchical data structures of the Intiendo address matching tool and described how they are implemented and used for address matching. We illustrated that Intiendo is an address-specific implementation of the ISO 19112 general model. Intiendo’s hierarchical fine tuning (setting an anchor level and to prioritizing Intiendo hierarchy levels for an address match) improves address matching considerably, and Intiendo’s *SpatialAdjacencyMatch* considers the spatial adjacency of potential address matches before suggesting a potential match, thereby eliminating irrelevant alphanumerically similar but spatially distant address matches. The implementation of a kd-tree for each location type in Intiendo is an extension of the data model described in ISO 19112 and improves the quality and performance of search results, especially for the spatial adjacent search. The approximations and filtering described above ensure that the spatial match is performed within acceptable response times.

Acknowledgements

This work is supported in part by the South African Department of Trade and Industry (dti) and AfriGIS (Pty) Ltd. The Intiendo address matching toolset is developed and distributed by AfriGIS.

References

Coetzee S and Cooper AK, ‘What is an address in South Africa?’, *South African Journal of Science (SAJS)*, Nov/Dec 2007, vol. 103, no. 11/12, pp449-458.

Davis AD Jr and Fonseca FT, 2007, ‘Assessing the certainty of locations produced by an address geocoding system’, *Geoinformatica*, **11**:103-129.

Goldberg DW, Wilson JP and Knoblock CA, 2007, 'From text to geographic coordinates: The current state of geocoding', *Journal of the Urban and Regional Information Systems Association (URISA)*, vol. 19, no. 1, pp33-46.

ISO 19111:2007 *Geographic information – Spatial referencing by coordinates*, 2003, International Organization for Standardization (ISO), Geneva, Switzerland.

ISO 19112:2003 Geographic information – Spatial referencing by geographic identifiers, 2003, International Organization for Standardization (ISO), Geneva, Switzerland.

SANS/CD 1883-1. Geographic Information – Address Standard, Part 1: Data format of addresses (committee draft), 2008, South African Bureau of Standards (SABS), Pretoria, South Africa.