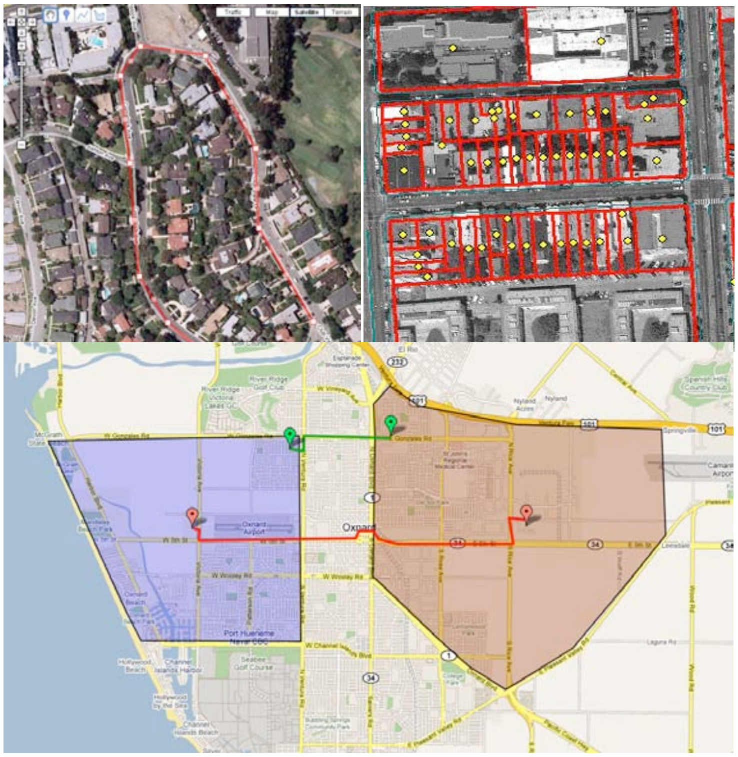
# Geocoding Best Practices: Reference Data, Input Data and Feature Matching



Daniel W. Goldberg
Jennifer N. Swift
John P. Wilson

Prepared for: Division of Cancer Prevention and Control
National Center for Chronic Disease Prevention and Health Promotion
Center for Disease Control and Prevention

**Cover Photos:**
Parcel boundaries (polygons) with centroids (points) (Google 2008), satellite imagery and vector base maps overlaid by streets (lines), and ZIP code boundaries (polygons) and address locations (point markers) (Los Angeles County Assessor 2008).

**Preferred Citation:**
Goldberg DW, Swift JN and Wilson JP 2008 Geocoding Best Practices: Reference Data, Input Data and Feature Matching. Los Angeles, CA, University of Southern California GIS Research Laboratory Technical Report No 8.

# Table of Contents

List of Figures ..... iv
List of Tables ..... v
Executive Summary ..... vi
1 Introduction ..... 1
2 National and Local Reference Datasets ..... 2
2.1 Linear Reference Datasets ..... 3
2.2 Polygon Reference Datasets ..... 5
2.3 Point Reference Datasets ..... 7
2.4 Reference Dataset Relationships and Accuracy ..... 8
2.5 Reference Dataset Best Practices ..... 9
3 Input Data Standards and Rules ..... 10
3.1 Address Styles ..... 12
3.1.1 Types of Input Data ..... 13
3.1.2 Handling Input Data in Different Resolutions ..... 15
3.1.3 Recommended Address Submission Formats ..... 17
3.1.4 Input Data Best Practices ..... 19
3.2 Address Cleaning ..... 20
3.2.1 Address Validation ..... 20
3.2.2 Address Validation Best Practices ..... 21
3.3 Address Normalization ..... 21
3.3.1 Substitution-Based Normalization ..... 21
3.3.2 Context-Based Normalization ..... 22
3.3.3 Probability-Based Normalization ..... 23
3.3.4 Address Normalization Best Practices ..... 24
3.4 Address Standardization Process ..... 25
3.5 Address Standardization Best Practices ..... 26
4 Feature Matching ..... 26
4.1 Feature Matching Best Practices ..... 28
4.2 Trouble-Shooting Solutions for Dealing with Non-Matches ..... 29
4.3 Trouble-Shooting Best Practices ..... 32
5 Conclusions ..... 33
6 Acknowledgements ..... 34
7 References ..... 35
8 List of Terms ..... 38

## List of Figures

1 High and low resolution examples of vector (line) reference data ........ 3
2 Example parcel boundaries (red) with centroids (yellow) (Goldberg 2008) ........ 7
3 Overview of the geocoding process (Goldberg 2008) ........ 11
4 A schematic of the generalized workflow in the geocoding process (Goldberg 2008) ........ 12
5 Example of a route determined using postal addresses (green) versus ZIP code centroids ........ 16
6 Generalized feature matching algorithm (Goldberg 2008) ........ 27

## List of Tables

1 Commonly used linear reference datasets (Goldberg 2008) .................. 4
2 Typical attributes for the linear reference datasets listed in Table 1 (Goldberg 2008) .................. 4
3 Commonly used polygon reference datasets (Goldberg 2008).................. 6
4 Attributes of widely used polygon reference datasets listed in Table 3 (Goldberg 2008) .................. 7
5 Commonly used point reference datasets (Goldberg 2008).................. 8
6 Point reference dataset attributes for the datasets listed in Table 5 (Goldberg 2008) .................. 8
7 Common types of input data, NAACCR fields and examples (Goldberg 2008).................. 14
8 Estimates of accuracy with respect to data resolution (Goldberg 2008).................. 16
9 Resolutions, issues, and ranks assigned to different types of addresses (Goldberg 2008) .................. 18
10 Example postal addresses in different formats (Goldberg 2008).................. 20
11 Common attributes of a city-style postal address (Goldberg 2008) .................. 22
12 Backus-Naur form of a postal address (Goldberg 2008).................. 23
13 Existing and proposed address standards (Goldberg 2008) .................. 25

## Executive Summary

The purpose of this report is to advise the Division of Cancer and Prevention and Control (DCPC), Centers for Disease Control and Prevention (CDC), and others who want to learn about reference and input datasets as well as specific standards and rules that should be followed in ensure the highest level of data cleaning, normalization, standardization and feature matching in their geocoding practices. This report begins with a discussion of national and local reference datasets typically used for geocoding cancer datasets. Reference datasets comprise the underlying geographic database that contains geographic features required by a geocoder to generate a geographic output. Though reference datasets are available in many forms and formats, this document focuses on the three types of reference datasets commonly used in geocoding processes: points, lines and polygons. The actual reference datasets most commonly used in geocoding cancer datasets for epidemiological research today are described herein. The geocoding best practices related to reference datasets are also identified. This is the first in a series of three reports which documents geocoding best practices for DCPC and CDC.

Next, cancer data input postal address styles (address style locator libraries) are described in terms of the different types of input data, how to handle input data of varying resolutions, and the recommended submission format or "gold standard" for input data. The process of address cleaning is explained along with the processes of address normalization and standardization. Address validation, the process of determining if an input address corresponds to a location that actually exists, is discussed in detail. The report also provides an overview of address normalization, the process of identifying the component parts of an address such that they may be transformed into some other desired format. Various normalization approaches ranging from the simplistic to advanced are defined, including substitution, context and probability-based normalization. To summarize these discussions, the best practices associated with cancer datasets including input postal address data, handling data in different resolutions, address cleaning/validation and the various methods of normalization are identified and listed at the end of each section. Address standardization, which is the conversion of an address from one normalized format into another, is also characterized and the address standardization best practices are summarized.

Lastly, this research report covers feature matching and trouble-shooting solutions for handling non-matches in the geocoding of cancer datasets. This discussion also includes detailed best practices with regards to feature matching and troubleshooting various issues that may arise during the feature matching process.

# 1 Introduction

Performing any type of research that involves geospatial mapping or investigation with the aid of a computer requires discrete, non-ambiguous, geographically valid digital data rather than descriptive text. In regards to cancer registries and epidemiological research, however, most data is reported in the form of postal addresses. This information typically includes the street address, city, and province or state of a patient at the diagnosis of their disease (Address, City, State). Although such text-based descriptions are easily understood by the layman, postal street addresses are not directly useable in a computerized environment. Georeferencing is the process of transforming textual information into geographically valid references that can be used for spatial analyses. Geographic information has the characteristics of volume, dimensionality, and continuity, such as geographic features that have the properties of size, distribution, pattern, continuity, neighborhood, shape, scale, and orientation (Clarke 2004).

Cancer-related health research and practice takes place across a multitude of administrative units and geographic extents. The reference and input data used to develop and address research questions are created, obtained, and processed by many different organizations with varying levels of expertise in data gathering and processing. Studies requiring the aggregation of data from multiple sources typically must integrate these disparate data, which often occur in incompatible formats with unknown lineage or accuracy. The inconsistencies and unknowns amongst these data can lead to uncertainty in the results that are generated if the data are not properly integrated. This problem of data integration represents a fundamental hurdle to cancer-related research. Thus a major goal of this and subsequent geocoding research papers is to offer advice on how best to handle these issues to ensure the highest level of confidence, reliability, standardization, and accuracy in geocoding activities.

It should be noted that the geocoding methods and reference data sources employed throughout all cancer registries in the United States (US) are quite diverse and varied (Abe and Stinchcomb 2008, Goldberg 2008). The North American Association of Central Cancer Registries (NAACCR) requires that all cancer registries adhere to address matching criteria by law, as described in NAACCR (2007). However, to date there are no legal mandates, organized directives or standards to guide the actual process of geocoding. A single definition explicitly defining, requiring, or endorsing a particular geocoding technology would not be useful. Within the cancer research community, each registry may have different restrictions or requirements on what information can be geocoded in terms of sources and types of input data (postal addresses, named places, etc.), which algorithms can be used for processing input, and what constitutes acceptable output. Differing levels of technical skills, varied access to geographic data, and budgetary and legal constraints can also require a broader definition of geocoding. As such, the definition offered herein is meant to serve the largest possible audience by specifically not limiting any of these characteristics of the geocoding process, intentionally leaving the door open for different definitions of geocoding to be considered valid.

Therefore, for the purposes of this report, the subsequent definitions are provided as background for the discussions to follow on reference datasets, input data standards and rules, and solutions for trouble-shooting and handling data in different resolutions. According to Goldberg (2008), “geocoding” is explicitly defined as the act of transforming descriptive locational text into a valid spatial representation using a predefined process. One of the goals of the next phase of this research is to generate recommendations on the standardization of the predefined process that will best serve the cancer registries and epidemiological research community. The term “geocoder” refers to a set of inter-

inter-related components in the form of operations, algorithms, and data sources that work together to produce a spatial representation for descriptive locational references. Geocoders are also referred to as "address locators" by ESRI (2008). The noun "geocode" is considered a spatial representation of descriptive location references, while "to geocode", the verb, is defined as the actual performance of the process of geocoding. These definitions help to resolve four common points of confusion about geocoding that are often complicated by disparate understandings of the term: the types of data that can be geocoded, the methods that can be employed to geocode data, the forms and formats of the outputs, and the data sources and methods that are relevant to the process. These definitions are intended to be broad enough to meet the diverse needs of both the cancer registry and cancer research communities (Goldberg 2008).

The primary purpose of this particular research report is to provide a set of recommendations or best practices for the process of geocoding for the cancer research community, which may also be used in the development of standardized methods for collection and transmission of geographic variables (e.g. addresses) by central cancer registries. These best practices will be framed as rules reflecting both policy and technical decisions that must be made by a registry as a whole as well as by the individual performing the geocoding or using the results. This report will cover the fundamental components of the geocoding process, including reference and input data, the internal processing performed, trouble-shooting recommendations, and how to handle disparate data in different geographic resolutions. For each component, choices that affect the accuracy of the resulting data will be presented and possible options that can be chosen will be listed. If these best practices are followed by the cancer registry community, the end result will be the establishment of a standardized knowledge base that will enable informed decisions within local registries, as well as the generation of consistent data that can be shared between organizations.

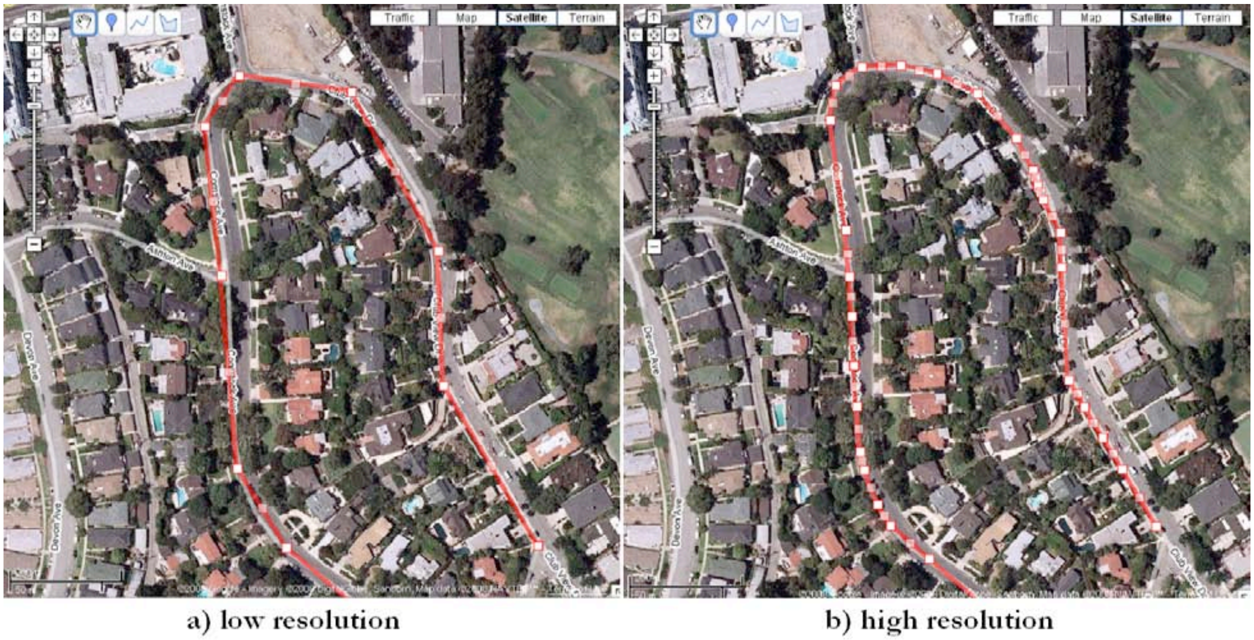
## 2 National and Local Reference Datasets

The reference dataset is the underlying geographic database containing geographic features that the geocoder uses to generate a geographic output. Reference datasets store all the information the geocoder knows about the world, and provide the base data from which the geocoder calculates, derives, or obtains geocodes. Interpolation algorithms use attributes of the input address to perform computations on the features contained in reference datasets, to estimate where the output of the geocoding process should be placed. Interpolation in geocoding practice is described in detail in Goldberg (2008). The sources of reference datasets can vary greatly from local government agencies, e.g. tax assessor offices, to national governmental organizations, e.g. US Census Bureau (2008a). Each must ultimately contain valid spatial geographic representations that can either be returned directly in response to a geocoder query (as the output) or be used by other components of the geocoding process to deduce or derive the spatial output (through interpolation).

Reference datasets are available in many forms and formats. There are three types of reference datasets used in geocoding processes: points, lines and polygons. The actual reference datasets most commonly used in geocoding practice today are described in this section, and are listed in Tables 1 through 6. Vector-based data are the most frequently encountered reference datasets in geocoding practice, because their per-feature representations allow for easy feature-by-feature manipulation. Features (geographic objects) are created by assigning properties to them that describe their location in space. Vector-based data are data composed of vector features, which are point-, line-, or area- (polygon)-based representations of data typically used to encode a range of values for a specific set of attributes for a single geographic feature. In contrast, raster-based data are composed of raster features, which are pixel-based representations of data typically used to encode a range of values for a specific set of attributes across a whole region. Raster-based data, such as digital orthophotos, can be harder to work with which makes them less applicable to geocoding. Nevertheless it should be noted that new geocoding processes are being developed that utilize raster-based data for tasks such as feature extraction and correction.

## 2.1 Linear Reference Datasets

A linear- or line-based reference dataset is composed of either simple lines or polyline (multiple line) vectors. The type of line vector comprising a dataset can sometimes be used as a first order estimate of the descriptive quality of the reference data source. Reference datasets containing only simple straight-line vectors will usually be less accurate than reference datasets containing polyline vectors for the same area, for instance if considering the shortest possible distance between two endpoints. Breaking single straight-line vectors into multiple segments is the usual way to represent curves and makes it possible to increase accuracy and resolution (Figure 1).



a) low resolution b) high resolution

**Figure 1 High and low resolution examples of vector (line) reference data**

Line-based datasets are by far the most cited in the geocoding literature, and are usually representations of street networks or graphs. The term network is defined herein as the topological connectivity resulting from reference features sharing common endpoints, such that it is possible to traverse through the network from feature to feature. Several well known examples of line-based reference datasets (street networks) are provided in Table 1. All the reference datasets listed in Table 1 possess the attributes described in Table 2, which are typically required in geocoding processes to perform feature matching using linear-based reference datasets. It is relevant to note that most of these attributes correspond directly to postal address-based input data.

### Table 1 Commonly used linear reference datasets (Goldberg 2008)

| Name | Description | Coverage | Cost |
|---|---|---|---|
| US Census Bureau's TIGER/Line | Street centerlines | US | Free |
| NAVTEQ Streets (NAVTEQ 2008) | Street centerlines | Worldwide | Cost varies according to coverage and user |
| TeleAtlas Dynamap, MultiNet (Tele Atlas 2008a, b) | Street centerlines | Worldwide | Cost varies according to coverage and user |

### Table 2 Typical attributes for the linear reference datasets listed in Table 1 (Goldberg 2008)

| Attribute | Description |
|---|---|
| Left side street start address number | Beginning of the address range for left side of the street segment |
| Right side street start address number | Beginning of the address range for right side of the street segment |
| Left side street end address number | End of the address range for left side of the street segment |
| Right side street end address number | End of the address range for right side of the street segment |
| Street prefix directional | Street directional indicator |
| Feature class code | A code representing the Census class of the feature (e.g. FCC) |
| Street name | Name of street |
| Street type | Type of street |
| Right side ZCTA | ZCTA for addresses on right side of street |
| Left side ZCTA | ZCTA for addresses on left side of street |
| Right side municipality code | A code representing the municipality for the right side |
| Left side municipality code | A code representing the municipality for the left side |
| Right side county code | A code representing the county for the right side |
| Left side county code | A code representing the county for the left side |
| Feature class code | A code representing the class of the feature |

Today, the US Census Bureau's TIGER/Line files are the most commonly used reference dataset in geocoding. The other two datasets in Table 1 are (or were at one time) commercial derivatives of the TIGER/Line files. All three products provide the same type of data, though the commercial versions contain more attributes and improvements over the TIGER/Line files in terms of reference feature spatial accuracy. The differences in the accuracy and cost between these products can be substantial. The commercial companies increasingly incorporate Global Positioning System (GPS)-level accuracy for their street network representations. Other geographic features such as hospitals, parks, and water bodies are often included along with network data that the company purchased or collected themselves. To cover the costs of such data collection tasks commercial data are usually very expensive, and may cost tens of thousands of dollars per state. Nevertheless the purchase price usually includes yearly or quarterly updates to the entire reference dataset, providing temporally accurate reference data.

Past releases of the TIGER/Line files have corresponded to the decennial Census, so that this source had temporal accuracy far behind their commercial counterparts. Also, most features are simple lines with very few other types of geographic features included. Nonetheless, while the commercial versions are very expensive, TIGER/Line files are free so they are still an attractive option. Some states and municipalities have created much higher-quality line files which will eventually be or have been already incorporated into the TIGER/Line files. Beginning in 2007 the US Census Bureau released MAF/TIGER files to replace annual TIGER/Line files (US Census Bureau 2008b). MAF/TIGER files merge the US Census Bureau's Master Address File, creating a relational database management system (RDBMS). A recent study by Ward et al. (2005) showed that in some areas the TIGER/Line files are as accurate as commercial files, and improving over time. Some of this change is due to the US Census Bureau's MAF-TIGER/Line file integration and adoption of the new American Community Survey system (US Census Bureau 2007a), which itself includes a large effort focused on improving the TIGER/Line files. There is also pressure from the US Federal Geographic Data Committee (FGDC) to improve TIGER/Line data (US Federal Geographic Data Committee 2008a). These advances are enabling greater public participation and facilitating the use of local-scale knowledge with higher accuracy of street features and associated attributes, e.g. address ranges, to inform and improve the national-scale products. Describing the relative accuracy within a given type of reference dataset (e.g. TIGER/Line) with respect to coverage and the effect on standardization of the geocoding process is a larger question which can be addressed in a subsequent phase of this research. In this context, a geocoder could be developed that provides metadata describing the accuracy of the reference datasets from which results are derived.

All linear features in these reference datasets typically include an attribute identifying the class of each feature, e.g. a major highway with a separator, minor road, tunnel. These classifications serve many functions including allowing for different classes of roads to be included or excluded during the geocoding process, enabling first-order estimates of road widths to be assumed based on the class of road, typical number of lanes in that class, and typical lane width. In the TIGER/Line files, these classifications are referred to as a Feature Classification Code (FCC) (US Census Bureau 2008b). In more advanced commercial datasets, supplementary information such as one-way roads, toll roads, etc., are provided as binary true/false values for each attribute.

## 2.2 Polygon Reference Datasets

Polygon-based reference datasets are composed of polygon data. Polygon reference features are often difficult and expensive to create, but they nevertheless offer higher accuracy than line or point datasets in many cases. For instance, a dataset representing true building footprints can provide an extremely accurate data source when based on ground surveys. Such data typically enable the geocoding process to return a result with a high degree of accuracy. Automated geocoding results of higher quality than this are generally obtainable only through the use of three-dimensional building models. Building footprints derived from photos would be of lesser or unknown accuracy. Similarly, the dataset quickly becomes less accurate and thus less appealing when polygons represent larger geographic objects or extents such as cities or counties. Most polygon-based datasets only contain single polygon representations, though some include multiple ring polygons. Three dimensional reference datasets such as building models are founded on multi-polygon depictions. Table 3 provides some examples of prevalent polygon-based vector reference datasets, along with estimates of their coverage and relative cost.

Recently both building footprints and three dimensional polygons are being used more frequently in commercial mapping applications. For example, Microsoft Virtual Earth and Google Earth both have three-dimensional coverage for hundreds of cities worldwide (Microsoft 2008, Google 2008). Nonetheless these datasets are difficult and costly to build. Though it is still rare for building footprints to be available for every building in an entire city, more and more are becoming available all the time.

### Table 3 Commonly used polygon reference datasets (Goldberg 2008)

| Name | Description | Coverage | Cost |
|---|---|---|---|
| TeleAtlas, NAVTEQ | Building footprints, parcel footprints, 5-Digit ZIP codes (US) | Worldwide, but sparse | Expensive |
| County or Municipal Assessors | Building footprints, parcel footprints | US, but sparse | Relatively inexpensive but cost and coverage varies by jurisdiction |
| US Census Bureau | Census block groups, census tracts, ZCTA, MCD, Counties, States | US | Free |
| US Postal Service (USPS) | 5-Digit Postal ZIP codes (vendors estimate boundaries from Census ZIP code Tabulation Areas) | US | Free to relatively inexpensive but cost varies by coverage and age of data |

Parcel boundaries are much more readily available today than building footprints (Figure 2). Parcel databases offer legally binding descriptions of property boundaries. They are usually generated through surveying, which is the most accurate method, though they may also be derived from imagery or some other form of legacy data. It is important to note that legally binding is not necessarily equivalent to highly accurate in every case. Parcel data are created by local governments for taxation purposes, and some states even mandate their creation and dissemination to the general public at low cost, e.g. California (Lockyer 2005). In addition, the FGCD has launched an initiative to create a single national level parcel file for the entire US, to be completed within a few years (Stage and von Meyer 2005). However, land and buildings not subject to local taxation, such as public housing, state-owned residential buildings or residences on military bases, may be omitted from a given dataset. Presently the cost of a parcel dataset can vary dramatically from one locality to another, ranging from free, e.g. Sonoma County, CA (County of Sonoma 2008), to very expensive, e.g. $125,000 for the Grand Rapids, Michigan Metropolitan Area (Grand Valley Metropolitan Council 2008). Most parcel reference datasets possess the attributes listed in Table 4.

Parcel-based reference features are discrete, meaning that they typically describe a single real world geographic feature. Thus, a feature matching algorithm will usually either find an exact match or none at all. Spatial operations can be performed on parcels to produce new related data such as centroids depicting the geometric centers of polygons. One drawback is that the address associated with a parcel may be the mailing address of the owner, not the address associated with the physical location of the parcel, referred to as the “situs address”. The extent to which this occurs is a larger research topic, to be addressed in a subsequent phase of this work. In most counties, assessors are under no mandate to include the situs address of a parcel in their databases. Thus the accessibility of such accuracy information is uncertain.

Currently low resolution versions of polygon reference datasets are readily obtainable. For instance the US Census Bureau freely offers polygon boundaries and their associated centroids for minor civil divisions, counties, and states. Thought these datasets may be of such low resolutions to be unsuitable for use as output, they can be very valuable when utilized as the boundaries of spatial queries when a feature matching algorithm is searching for a linear reference feature within another reference dataset. For example, they can be used to constrain or clip the spatial area that must be searched, which could significantly speed up one or more data processing operation(s).

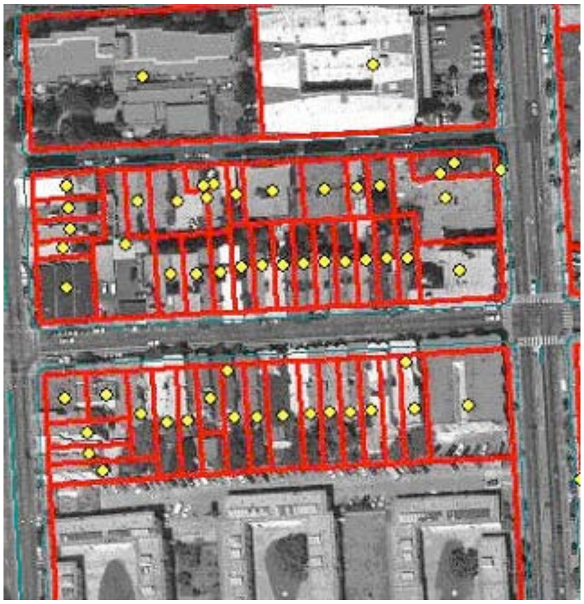


**Figure 2 Example parcel boundaries (red) with centroids (yellow) (Goldberg 2008)**

**Table 4 Attributes of widely used polygon reference datasets listed in Table 3 (Goldberg 2008)**

| Attribute | Description |
|---|---|
| Name | The name of the feature used for search |
| Polygon Coordinates | Set of polylines in some coordinate system |
| Index code/identifier | Code to identify the polygon within the reference data system |

## 2.3 Point Reference Datasets

A point-based reference dataset is composed of point-based data. Point-based datasets are the least commonly used reference datasets in geocoding practice due to their limited usability and varying cost and accuracy, although address point datasets are becoming more available. Using a point reference dataset in a geocoder will only return values for input addresses that actually exist (unless tables containing aliases are available), and though the match accuracy can be high, the match rate will be lower than when line or polygon reference datasets are utilized. And although linear and polygon datasets can handle values within ranges for a feature to be matched, precision will be lower than that obtained using point-based datasets. The cost of production and accuracy of point-based reference datasets can range from extremely high when using GPS devices to extremely low cost and variable accuracy when utilizing legacy geocoded data (generated from other sources). Several examples of frequently used national scale point reference datasets are provided in Table 5, and attributes they have in common are listed in Table 6. The costs of these datasets vary from free to expensive, depending on the customer as well as the amount of coverage. The latter comprise the minimum set of attributes required for a feature matching algorithm to successfully match a reference in a point-based reference dataset.

### Table 5 Commonly used point reference datasets (Goldberg 2008)

| Name | Description | Coverage |
|---|---|---|
| E-911 Address Points | Emergency management points for addresses | Portions of US |
| Postal Codes | Postal Code centroids | US |
| Census MCD | Minor Civil Division centroids | US |
| GNIS (US Board on Geographic Names 2008) | Gazetteer of geographic features | US |
| GeoNames (US National Geospatial-Intelligence Agency 2008) | Gazetteer of geographic features | World, excepting US |
| ADL (ADL 2008) | Gazetteer of geographic features | World |

### Table 6 Point reference dataset attributes for the datasets listed in Table 5 (Goldberg 2008)

| Attribute | Description |
|---|---|
| Name | The name of the feature used for the search |
| Point coordinates | A pair of values for the point in some coordinate system |

## 2.4 Reference Dataset Relationships and Accuracy

Unfortunately, the US does not currently possess a national-scale reference dataset containing accurate geocodes for all addresses in the country, though some are available at local scales. The national-scale datasets contain low resolution geographic features and are mostly available from the US Census Bureau. One example is ZIP code Tabulation Area (ZCTA) centroids and points representing named places such as minor civil divisions, which are distributed along with TIGER/Line files (US Census Bureau 2008b). ZIP codes are different than ZCTA, and their (approximate) centroids are available from the USPS and numerous commercial vendors (Rushton et al. 2008). Although higher resolution point data have been generated by individual localities in the US, these can be difficult to obtain unless one is active or has connections in the locality of interest. The establishment of a national geospatial database containing ZIP codes and ZCTA would be a help to geocoding practice.

It is important to consider the relationships that exist between different reference dataset types because they can be structured in different ways, as spatially-hierarchical relationships and lineage-based relationships. An example of the first is the relationship between polygon-based features available at different geographic resolutions of Census delineations in the TIGER/Line files. Census block groups offer the highest resolution, followed by census tracts, ZCTA, county subdivisions, counties, and/or other state subdivisions. In many cases data at lower resolutions represent an aggregation of the features at the higher level. When choosing a reference feature for interpolation, one can safely change from selecting a higher resolution representation to a lower one, e.g. a censusa block group to tract, without fear of introducing erroneous data. Since lower resolution data are composed of multiple higher resolution features, the reverse is not true. A level of ambiguity will be introduced as to which higher resolution feature to select if attempting to increase the resolution of the feature type matched to.

In addition to the relationships among different datasets, some other relationships between features within a single dataset also require attention. These include holistic and atomic metrics, which are used to describe datasets. The holistic metrics refer to characteristics that describe values over an entire dataset, whereas atomic metrics describe individual features in a dataset. For example, TIGER/Line files claim the holistic metric "average horizontal spatial accuracy" as a single value, e.g. 7 m. An example of an atomic metric would be the accuracy of individual polygons within a given dataset. Another characteristic related to atomic and holistic feature completeness and accuracy is referred to as geographical bias, which is defined herein as the observation that the accuracy of geographic features may be a function of the area where they are located (Boscoe 2008).

The spatial accuracy of the reference datasets used by a geocoding process may be the most critical factor contributing to the spatial accuracy of the output, since different representations of reference features encode different levels of information and can thus be highly variable. Interpolation algorithms operating on the reference features can only work with what they are given, and will never produce any result more accurate than the original reference feature. Though interpolation algorithms can and do produce spatial outputs of varying degrees of spatial accuracy based on their intrinsic characteristics, the baseline accuracy of the reference feature is translated directly to the output of the interpolation algorithm. It is sometimes true that the larger the geographic coverage of a reference dataset, the worse the spatial accuracy of its features. For instance this has been the case historically when comparing street vectors based on TIGER/Line files, to those produced by local governments. For more detailed information on spatial accuracy and quality in regards to reference datasets, please see Section 13 in Goldberg (2008).

## 2.5 Reference Dataset Best Practices

To summarize the discussion in this section, the following list is a compilation of the overarching best practices related to reference datasets:

- Any reference dataset format should be supported by a geocoding process: linear, polygon or point-based, vector and raster datasets. At a minimum, vector-based and/or linear-based datasets must be supported.
- A cancer registry should obtain the most accurate reference dataset they can given their budgetary and technical constraints. Cost may be the deciding factor as to which data source to use. There may be per-product limitations, so all choices should be fully investigated before acquisition.
- New data sources should be obtained regularly in order for a cancer registry to keep their

reference dataset up-to-date. The update frequency will depend on budgetary constraints and the frequency with which vendors provide updates.

- Old data should not be discarded. A cancer registry should retain historical versions of all their reference datasets.
- In many cases reference datasets can be obtained from local government agencies. The FGDC should also be contacted to determine the types, numbers, and usability of reference datasets available. Commercial firms (e.g. Tele Atlas and NAVTEQ) can also be contacted if needs cannot be met by public domain data.
- Cancer registries should maintain lists of reference datasets applicable to their area across all resolutions (e.g. TIGER/Line - national, county roads - regional, parcel databases - local).
- In regards to reference datasets, researchers and/or staff should be trained on how to maintain as well as utilize the datasets.
- Primary source reference datasets should be preferred to secondary derivatives unless significant improvements have been made and are fully documented and can be proven.
- If the geographic variability of a region is low or the size of the region covered is small (e.g. city scale), the holistic metrics for the reference dataset should be used. Conversely, if the geographic variability of a region is high or the size of the region covered is large (e.g. national scale), the accuracy of individual reference features within the area of the input data should be considered over the holistic measures, if available.
- Estimates of the atomic feature accuracy within a reference dataset should be made periodically by random selection and manual evaluation of the reference features within the region covered by the dataset.
- Geographic bias should be considered a potential problem if the geographic variability of a region is high or the size of the region covered is large (e.g. national scale).

# 3 Input Data Standards and Rules

Looking beyond the reference dataset(s) used, the geocoding results will be significantly impacted by the input data standards and rules. The technical background and input data requirements that makes the geocoding process possible are discussed in this section. To begin, a generalized illustration of the geocoding process is provided in Figure 3. This process involves three separate yet related components: the descriptive locational input data, the geocoder, and the spatial output data (Goldberg 2008). The interactions between the different components illustrate the basic steps that a typical geocoder performs as it produces output from the input provided to it. The input data to the geocoding process is understood to be any descriptive locational textual information such as an address or building name. A geocoder is composed of two fundamental components, the reference dataset and geocoding algorithm, each of which may be composed of a series of sub-components and operations. The output data can be any form of valid spatial data (e.g. latitude and longitude). Geocoding is considered the actual process used to convert the input into the output. These individual steps, related issues and best practices will be described in more detail in the subsections that follow.

The actual software implementation of a geocoder will vary in terms of the components chosen and conceptual representation of the geocoding system, depending on the user. Each cancer researcher will have his/her own geocoding requirements, or set of constraints that affect the choice of geocoding options. Technical, budgetary, legal, and policy constraints will all influence the choice of geocoding requirements. These requirements should be reviewed, updated and/or changed annually, at a minimum. Even though the geocoding requirements may vary within the cancer research community, standards for data reporting spatial fields are available from NAACCR, and from the organization Health Level Seven, one of several American National Standards Institute (ANSI) accredited health research focused standards developing organizations (Health Level Seven 2007, Hofferkamp and Havener 2008). The NAACCR standards reference United States Postal Service *Publication 28 - Postal Addressing Standards* (US Postal Service 2008d).

Figure 4 illustrates the general workflow of the geocoding process in more detail. The essential components that should be common to any geocoder implementation are included. It also shows the decisions that may need to be made by cancer researchers, depending on their requirements. This diagram also illustrates the various steps and requirements that geocoder vendors will need to accommodate in order to work with the cancer research community.

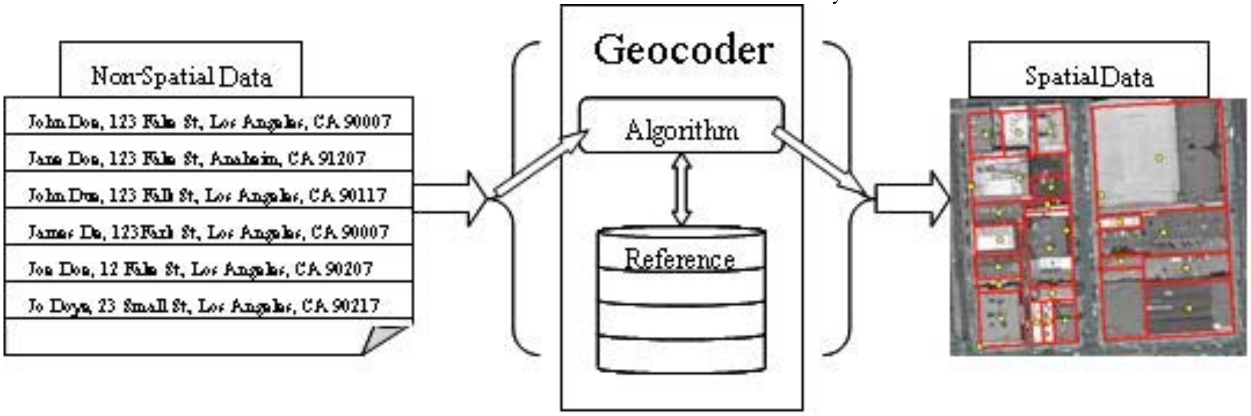

Non-Spatial Data
John Doe, 123 Fake St, Los Angeles, CA 90007
Jane Doe, 123 Fake St, Anaheim, CA 91207
John Dwe, 123 Fall St, Los Angeles, CA 90117
James De, 123 Fark St, Los Angeles, CA 90007
Joe Doe, 12 Fake St, Los Angeles, CA 90207
Jo Doye, 23 Small St, Los Angeles, CA 90217
Geocoder
Algorithm
Reference
Spatial Data

**Figure 3 Overview of the geocoding process (Goldberg 2008)**

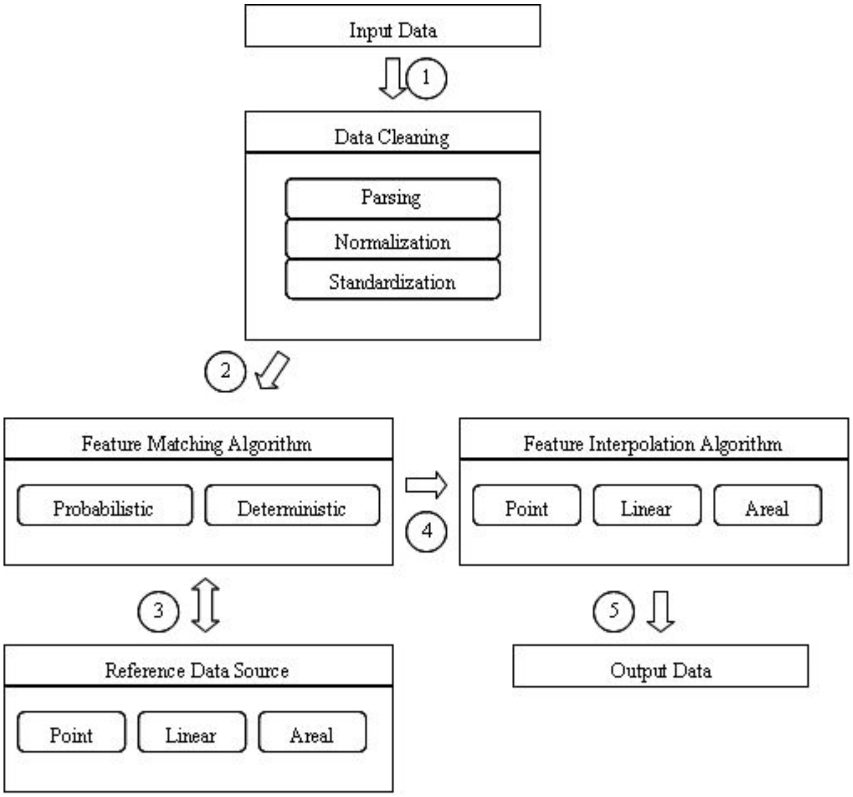


**Figure 4 A schematic of the generalized workflow in the geocoding process (Goldberg 2008)**

## 3.1 Address Styles

For the purposes of this report, input data are defined herein as the descriptive locational text that are to be converted into computer useable spatial data through the process of geocoding (Goldberg 2008). Table 7 provides examples of the wide variety of possible forms and formats of input data. On the one hand this diversity demonstrates the flexibility which should be inherent in a geocoder, and on the other it illustrates how the multiplicity of options is a contributing factor to the overall difficulty of implementing geocoding. The recommended address submission format for cancer registries is provided in section 3.1.3.

Input data can be initially classified into two categories, referred to as relative and absolute. Relative input data are defined as textual location descriptions which produce relative geocodes, geographic locations that are relative to some other reference geographic locations. Alone, these are not sufficient to produce an output geographic location. An example would be an interpolated distance along or within a reference feature, as in the case of lines and polygons, respectively. Without a reference to geographic locations such as a line vector or polygon, it is impossible to obtain output locations. An example of relative input data would be "Across the street from city hall". These are not typically considered valid address data for submission to a cancer registry, but they are often submitted anyway. In the future, each cancer registry could identify which submissions are classified as absolute versus those classified as relative, if desired. Many geocoding platforms do not support relative input and thus they will not be matchable. Absolute input data are defined as textual location descriptions which on their own can produce an output geographic location. Absolute input data produce absolute geocodes in the form of a known location, or an offset from an absolute known location. Adequately referenced place names, ZIP codes, or parcel identifiers are examples of absolute input data. These can be directly looked up in an available data source to obtain an output geocode. Linear addressing system locations are also absolute by definition. For example, the Emergency 911-based (E-911) geocoding systems are absolute because they use distances from known mileposts on streets as coordinates. Mileposts represent a linear addressing system, assuming each milepost is an absolute known location.

### 3.1.1 Types of Input Data

The most common form of input data encountered in cancer related research is the postal address (Goldberg 2008). Postal addresses come in many different forms, including those shown in Table 7. A city-style postal address describes a location in terms of a numbered building along a street. A Rural Route (RR) or Highway Contract (HC) address is intended to identify a stop on a postal delivery route. And, a USPS Post Office (PO) Box address represents a physical storage location at a US Post Office or some other mail handling facility. Examples of each form are provided in Table 7.

A city-style postal address can identify locations down to sub-parcel and floor-plan levels. The attributes of a city-style postal address generally include a house number and street name, along with a city, state, and USPS ZIP code. In addition, each attribute may be broken down further into more descriptive levels, such as unit numbers, fractional addresses, and USPS ZIP+5 codes (US Postal Service 2008a). Pre- and post-directional attributes are used to differentiate individual streets when several in the same city have the same name and are within the same USPS ZIP code, such as when the origin of the address range of a street is in the center of a city and expands outward in opposite directions. Since city-style postal addresses are so common, suitable reference datasets and geocoders capable of processing it are widely available at many different levels of accuracy, resolution, and cost (see Tables 1 and 3). Nevertheless several drawbacks exist when city-style postal addresses are used, due to the large number of possible attributes that grants these addresses their descriptive power. Significant problems can occur during feature matching if attributes are missing, not ordered correctly, or if extraneous information has been included. Other issues include when the same values are used for multiple attributes, e.g. directional indicators like “400 East West Avenue”, and if non-English-based attributes are used. Another serious problem arises due to a class of locations which have ordinary city-style postal addresses, but do not receive postal delivery service. An example of this is a private development or gated community. These data may sometimes be the most difficult cases to geocode because postal address-based reference data are truly not defined for them, and systems relying heavily on postal address-based normalization or standardization may fail to process them.

### Table 7 Common types of input data, NAACCR fields and examples (Goldberg 2008)

| Data Type | NAACCR Field(s) | Example |
|---|---|---|
| Complete postal address | 2330: dxAddress - Number and Street<br>70: dxAddress - City<br>80: dxAddress - State<br>100: dxAddress - Postal Code | 1840 Century Park E, # 1200, Los Angeles, CA 90007-21000 |
| Partial postal address | 2330: dxAddress - Number and Street | 1840 Century Park |
| PO Box | 2330: dxAddress - Number and Street<br>70: dxAddress - City<br>80: dxAddress - State<br>100: dxAddress - Postal Code | PO Box 4567, East Bay, RI 08040-2309 |
| Rural Route | 2330: dxAddress - Number and Street<br>70: dxAddress - City<br>80: dxAddress - State | RR 3 BOX 12000, Torrance CA |
| City | 70: dxAddress - City | Newport |
| County | 90: County at dx | San Luis Obispo County |
| State | 80: dxState - State | NY |
| ZIP code, ZIP+4 | 100: dxAddress - Postal Code | 12333-6789 |
| Intersection | 2330: dxAddress - Supplemental | Century Park E and Constellation Blvd |
| Named place | 2330: dxAddress - Supplemental | University of Southern California |
| Relative | 2330: dxAddress - Supplemental | Southeast corner of South Santa Monica Blvd and Century Park E |

A USPS PO Box address can only provide the USPS ZIP code attribute and does not represent the residences of individuals. In most cases USPS PO Box data cannot be geocoded to street level accuracy. One example would be the situation where a person rents a USPS PO Box at a Post Office near their work, but lives in a completely different city. In the same manner personal mail boxes may be reported and have the same lack of correlation with residence location. Since this is a significant problem in geocoding practice, a considerable amount of research has been dedicated to the effect of USPS PO Boxes on geocoding processes (e.g. Hurley et al. 2003, Shi 2007). In some cases geocoding using USPS PO Box delivery-weighted five-digit ZIP code centroids is utilized as a proxy for street address, though the misclassification of boxholders can be significant (Hurley et al. 2003). Besides ZIP centroid, attribute imputation based on other characteristics known about the person or the area can be applied (e.g. Boscoe 2008), or a distribution function can be used to place the geocode in a specific location (geocode imputation/pseudocode) (Rushton et al. 2008). The development of recommendations for geocoding USPS PO Boxes is a sizeable research topic unto itself, which could be addressed in the next phase of this project.

An RR and/or Highway Contract (HC) address is most often found in rural areas and is written "RR 21 Box 5", which indicates that mail should be delivered to "Box 5" on the rural delivery route

"Number 21". Delivery locations can be a single mailbox at a single residence, or a physical cluster of several boxes at a single drop-off point where multiple residents pick up their mail. Historically, numerous problems have been documented when geocoding RRs and HCs, since they represent a route (path) traveled by a mail carrier rather than a single street, and, the Box number attribute does not include any data needed for feature interpolation. In addition, there is no information on whether a box is standalone or within a cluster, nor is it possible to estimate a relative distance along a reference feature. Thus, it was unquantifiable and unusable in an interpolation algorithm. Although in the past it was not possible to derive a single street name from a numbered RR portion of an RR address, advances are now being made to address this issue due to the continuing implementation of the E-911 service. To comply with E-911 regulations, local governments assign actual geocodes to the RR addresses (and their associated phone numbers) based on the existing linear-based referencing system of street mileposts. Hence a new system of absolute geocodes for RR addresses has been generated for many RR's across the US. Most significantly for geocoding, the USPS has created the Locatable Address Conversion System (LACS, US Postal Service 2008e) database which supports conversion of RR to city-style postal street addresses (US Postal Service 2008c), which supports and provides a direct link between an RR postal address and the reference datasets capable of interpolation-based geocoding which require city-style postal addresses (where E-911 has been implemented). The USPS has mandated that all Coding Accuracy Support System (CASS) Certified software providers must support the LACS database to remain certified (US Postal Service 2008b), so RR to city-style address translation is available now for a modest fee for most areas in the US.

It is important to note that the common misunderstanding between USPS ZIP codes and US Census Bureau ZCTAs is that the two refer to the same thing and can be used interchangeably, despite the fact that their differences have been widely published (Rushton et al. 2008). USPS ZIP codes represent delivery routes rather than regions, while a ZCTA truly represents a contiguous geographic area. The resulting negative effects on the geocoding process have been widely publicized and documented in the geocoding literature (e.g. Krieger et al. 2002, Hurley et al. 2003, Grubesic and Matisziw 2006, Beyer et al. 2008).

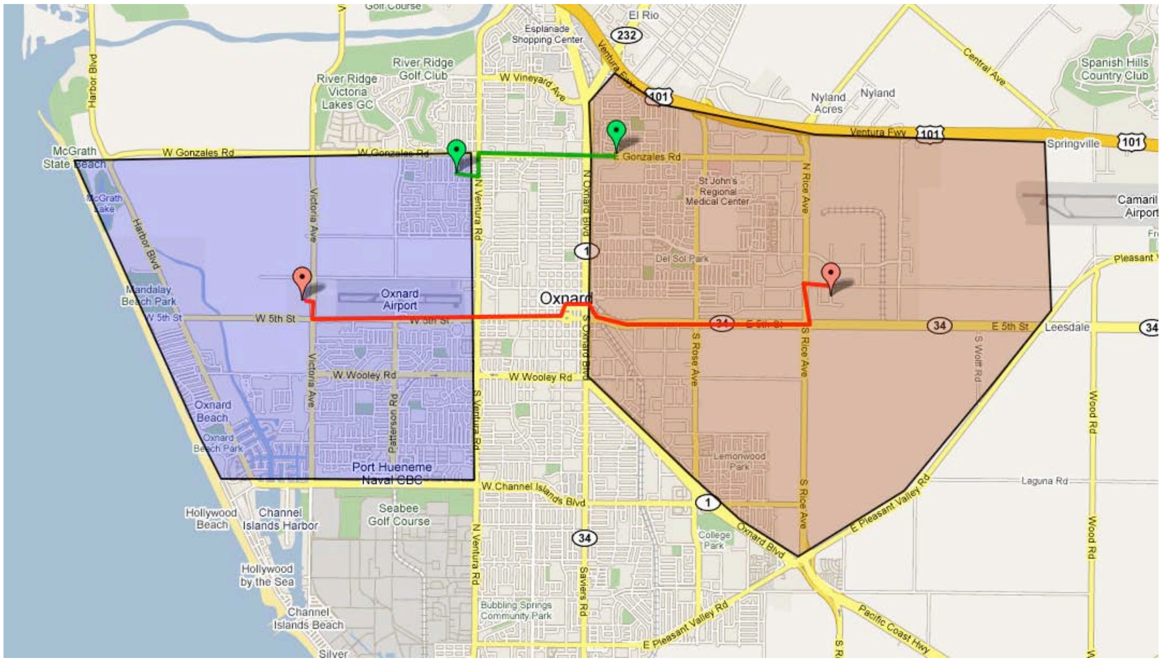
From the list of possible input data formats in Table 7, it can seen that most input data are based on postal addressing systems, administrative units, named places, coordinate systems, or relative descriptions that use one of the others as a referent. Input data in the form of complete or partial city-style postal addresses are most often encountered. As stated previously, significant problems may appear when processing postal address data both because of the high degree of variability in the way they can be represented, and the fact that they often include extraneous data and/or are missing required elements. In order to address these issues geocoders may utilize data processing methods such as address normalization and address standardization. These methods are discussed in Sections 3.3 and 3.4, respectively.

### 3.1.2 Handling Input Data in Different Resolutions

Due to the different possible resolutions of the various types of input address data, best and worst cases can be identified (Goldberg 2008). With respect to input data, resolution refers to the level or amount of information provided in a given input address. Table 8 provides the first order accuracy estimates one can expect to achieve in terms of geographic resolution, with respect to the input data types commonly used in geocoding practice. Although some research has been conducted on associating first order levels of accuracy with different types of location descriptions (e.g. Davis Jr. et al. 2003, Davis Jr. and Fonseca 2007), in actual practice these distinctions are seldom quantified and returned as accuracy metrics with the resulting data. The bottom line is that different types of data descriptions (e.g. relative versus absolute) inherently contain different levels of information. Figure 5 illustrates what a dramatic difference using postal addresses versus ZIP code centroids as input data can have on the accuracy of routes generated from the these two types of  locational descriptions.

### Table 8 Estimates of accuracy with respect to data resolution (Goldberg 2008)

| Data Type | Best Case Scenario | Worst Case Scenario |
|---|---|---|
| Standard postal address | Sub-parcel | State |
| USPS PO Box | USPS ZIP code centroid | State |
| Rural Route | Sub-parcel | State |
| US National Grid | 1 m$^2$ | 1000 m$^2$ |



### Figure 5 Example of a route determined using postal addresses (green) versus ZIP code centroids

City-style postal addresses are exceptionally useful since the information they contain is hierarchical in structure. This embedded hierarchy is often used as the basis for multi-resolution geocoding processes that allow varying levels of geographic resolution in the resulting geocodes based upon where a match can be made in the hierarchy. An example of a city-style postal address which has all possible attributes filled in (excluding multiple street type suffixes) is provided in the first row of Table 9. This table illustrates the sequence of geographic resolution in terms of different combinations of address attributes. The possible variations of this address are also listed in Table 9, ranked from lowest (8) to highest (0) in order of decreasing geographic resolution. The best possible and most probable resolutions are also shown, and also the uncertainty introduced at each resolution. It is important to note that eliminating attributes from city-style postal addresses quickly degrades the best possible accuracy, and that different combinations of attributes will have a significant impact on the geographic resolution or granularity of the resulting geocode. For a more detailed discussion on the strengths and weaknesses of arbitrarily ranking geographic resolutions, see Goldberg (2008).

### 3.1.3 Recommended Address Submission Formats

The address example in the first row of Table 9 illustrates the "gold-standard" in postal address data (Goldberg 2008). This example contains valid information in each of the possible address attribute fields and indicates enough information to produce a geocode down to the sub-parcel unit or the floor level. A hierarchical feature matching algorithm implements a geographic scale progression where a search for such an address is first confined by a state, then by a city, then by a detailed USPS ZIP code, in order to limit the number of possible features to a given area. Any ambiguity associated with the street name can be removed using prefix and suffix directionals associated with the name, in this case "South" and "East", respectively. In the next step, parcel identification can be performed using the street number, "1840", assuming that a parcel reference dataset exists and is accessible to the feature matching algorithm. Lastly, a three-dimensional geocode can be produced from the sub-parcel identification by combining the unit indicators "½" and "Unit 1200" to determine the floor and unit on the floor, assuming that this is an apartment building and a 3D building model is available to the feature matching algorithm. Note that both "½" and "1200" can mean different things in different localities, such as subdivided parcels or subdivisions within a given parcel.

This gold-standard address exemplifies the best case scenario with regards to postal address requirements and reference dataset availability. In most cancer-related data, gold-standard addresses are hardly ever obtained. This problem can be attributed to three factors: high quality reference datasets do not exist for many regions; details such as the floor plan within a building are rarely requested; and address input data are almost never this detailed. It is often assumed that utilization of the USPS ZIP+4 database will provide the gold-standard reference dataset, but it is actually only the most up-to-date source for address validation. The USPS ZIP-4 database must be used in conjunction with other sources or reference datasets to obtain the spatial aspect of an output geocode.

In practice, the form of address in the fourth row of Table 9 depicts the most frequently encountered address style. This example lacks the street directional, sub-parcel, and additional USPS ZIP code components of the address. A feature matching algorithm processing this case could quickly limit its search for matching reference features to within the USPS ZIP code. However, from this point problems may arise due to "address ambiguity", the case where a single input address can match more than one reference feature. This issue generally indicates an incompletely described input address, and can occur at multiple levels of geographic resolution for many different reasons. This is also an example of "street segment ambiguity", where multiple street segments could all be chosen as the reference feature for interpolation based on the information available in the input address. To begin with, multiple streets within the same USPS ZIP code can possess the same name, the only difference being in the directional information associated with them (e.g. which side of a city they are located on). Additionally, the address range information commonly associated with street reference features is frequently repeated for these streets. The end result is that the feature matching algorithm may be offered multiple options, each of which might be suitable as a match for the input address. Furthermore, "street address ambiguity" may occur on an even finer scale, where a single input address can match more than one reference address on a single street segment. In such a scenario, a correct street segment can unambiguously be determined, but a specific location along the street cannot because the address number is missing. The fifth row of Table 9 presents an example of such an address. Last of all, at the finest scale, "sub-parcel address ambiguity" can occur if a single input address can match more than one reference feature contained within the same parcel. This problem often occurs with large building complexes. In such ambiguous cases, most feature matching algorithms do not contain sufficient information to be able to associate the correct feature with the address. Details on feature matching procedures are presented in Section 4, and a summary of the different trouble-shooting methods for dealing with scenarios such as these is presented in Section 4.2.

**Table 9 Resolutions, issues, and ranks assigned to different types of addresses (Goldberg 2008)**

| Address | Best Resolution | Probable Resolution | Ambiguity | Rank |
|---|---|---|---|---|
| 1840 ½ South Century Park East, No. 1200, Los Angeles, CA 90007-21000 | 3D Sub-parcel-level | Sub-parcel-level | None | 0 |
| 1840 South Century Park East, Los Angeles, CA 90007-21000 | Parcel-level | Parcel-level | unit, floor | 1 |
| 1840 South Century Park East, Los Angeles, CA 90007 | Parcel-level | Parcel-level | unit, floor, USPS ZIP code | 2 |
| 1840 Century Park, Los Angeles, CA 90007 | Parcel-level | Street-level | unit, floor, street, USPS ZIP code | 3 |
| Century Park, Los Angeles, CA 90007 | Street-level | USPS ZIP code-level | building, unit, floor, street, USPS ZIP code | 4 |
| 90007 | USPS ZIP code-level | USPS ZIP code-level | building, unit, floor, street, city | 5 |
| Century Park, Los Angeles, CA | City-level | City-level, though small streets may fall entirely into a single USPS ZIP code | building, unit, floor, street, USPS ZIP code | 6 |
| Los Angeles, CA | City-level | City-level | building, unit, floor, street, USPS ZIP code | 7 |
| Century Park, CA | State-level | State-level | building, unit, floor, street, USPS ZIP code, city | 8 |
| CA | State-level | State-level | building, unit, floor, street, USPS ZIP code, city | 8 |

For the purposes of geocoding, incorrectly formatted addresses and those possessing non-standard abbreviations should be handled by the address normalization and standardization processes discussed in Sections 3.5, 3.6 and 3.7. In addition, manual methods may also be employed to normalize and standardize address. The geocoding best practices related to input data types, resolution and acceptable address submission formatting are listed in Section 3.2.1, and those corresponding specifically to address normalization and standardization are provided in Sections 3.5 and 3.7, respectively.

### 3.1.4 Input Data Best Practices

The best practices relating to the aforementioned input postal address data can be listed as follows:

- Any type of address data should be considered valid geocoding input (e.g. city-style and rural route postal addresses).
- Input data should be formatted as city-style postal addresses whenever possible.

- If possible, USPS PO Box data should be investigated to obtain more detailed information and formatted as city-style postal addresses to be considered acceptable for geocoding.
- If possible, RR and HC data should be converted into city-style postal addresses to be considered acceptable for geocoding.
- If possible, USPS ZIP code and/or ZCTA data should be investigated for more detailed information and formatted as a city-style postal address to be considered acceptable for geocoding. If USPS ZIP code and/or ZCTA data must be used, special care needs to be taken when using the resulting geocodes in research (Beyer et al. 2008).
- If the potential level of resulting accuracy is too low given the input data specification and the reference features that can be matched, lower level portions of the input data should be used (e.g. USPS ZIP code, city).
- If legitimate attributes of an address are missing and can be non-ambiguously identified, they should be added to the address to make it a "Gold Standard". Metadata should be created that includes which attributes were added, and which sources were used. The NAACCR minimum metadata standard for input data submitted to cancer registries must be followed. In addition, requirements could be amended to include standardized metadata about the geocoding process.
- Incorrect portions of input address data should be corrected if information is available to deduce the correct attributes. Metadata should be created that includes the information used in the selection, the attributes corrected, and the original values.
- If input address data are incorrectly formatted, and if the data is formatted in a known format, the address normalization process could be applied to try to identify the components of the address and subsequently reformat it into a more standard format, which should be noted in the metadata. However, if the format of the original data is unrecognizable or the address normalization fails, it should be left in its original format.
- If input address data that include non-standard abbreviations, address normalization and standardization components of the geocoding process should be applied to correct the data and the corrections should be noted in the metadata. However, if these processes fail, the data should be left in its original format.
- Any extraneous input address data or information describing the location or address should be moved into the supplemental field for retention in the case that it becomes useful in the future.
- At a minimum, researchers and/or staff should be trained on how input data should be formatted and corrected.

## 3.2 Address Cleaning

One of the most noteworthy contributing factors to the success or failure of a geocodeing process is called the "cleanliness" of the input data. Address data are often referred to as "dirty" due to human error (e.g. simple data entry mistakes) as well as the use of non-standard abbreviations and attribute orderings. For instance, the addresses cited in Table 10 are all in different formats yet refer to the same address, which illustrates why address cleaning is required. Addresses must first be cleaned in order to prepare input address data for geocoding. The address cleaning process relies on address normalization and/or address standardization. These address cleaning methods (also referred to as input data processing) and the best practices (rules) associated with each are described in the following subsections.

**Table 10 Example postal addresses in different formats (Goldberg 2008)**

| Examples |
|---|
| 1840 Century Park East, # 1200, Los Angeles, CA 90007-21000 |
| 1840 Century Park E, 1200, Los Angeles, CA, 90007-21000 |
| 1840 CENTURY PARK E, UNIT 1200, LA, CA |
| E Century Park 1840, Los Angeles, CA, 90007 |

### 3.2.1 Address Validation

Address validation is a key component of address cleaning. Address validation is the process of determining if an input address corresponds to a location that actually exists (Goldberg 2008). It is recommended that address validation always be attempted.  This has a direct effect on the accuracy of the geocode produced. Ideally, validation should occur at the origin of data gathering, such as a hospital, rather than at a secondary data source, e.g. at a cancer registry. The most commonly used source used for address validation is the USPS ZIP+4 database (US Postal Service 2008a), but other sources may be available for different localities and may provide additional information. US Census Bureau Census tracts and county or municipal assessor parcels are also frequently used to perform address validation. Note that even though some addresses may validate, they still may not be geocodable due to issues with the reference dataset and visa versa.

The simplest way to perform address validation is to perform feature matching using a reference dataset that contains discrete features. As previously mentioned in Section 2.2, discrete features such as points are single features that represent only single real world entity features, as opposed to a line or polygon feature which represents a sequence or range of real world entities. If feature matching applied to a reference dataset of discrete features is successful, the matched feature returned can be described as either "true positive" or "false positive". True positive is defined when an input address is returned as being true, and is in fact true, e.g.  it actually exists in the real world. Whereas false positive is defined as the case where an input address is returned as being true while in fact it is false, e.g. it does not really exist. If feature matching fails (see Section 4.2) the input address is usually specified as "true negative" or "false negative". The definition of true negative is the case when an input address is returned as being false, and it really is false. A false negative is defined as an input address which is returned as false, but it is actually true, e.g. it does exist in the real world. False positives and negatives can also occur due to temporal inaccuracy of reference datasets, for instance when the input address is invalid but appears in the reference dataset, or when the input address exists but has not yet been added to the reference dataset, respectively. Therefore, the level of confidence for the temporal accuracy of a reference dataset must be ascertained and utilized. The level of confidence can be assessed by considering the frequency of reference dataset update, address lifecycle management in the region, and characteristics of the region. This type of information can include the age of the reference set, how often updates occur, and how often addresses change in the region. More details on temporal accuracy in reference datasets can be found in Goldberg (2008).

Although parcel data have proven to be a very valuable source of address data, as previously stated it is noteworthy that in most counties in the US assessors are under no mandate to include the situs address of a parcel in their databases. In some cases the mailing address of the parcel owner may be the only information available, but it may or may not be the real address of the specified parcel. E911 address points provide an alternative for performing address validation. Current research on the effect of discrete (address point- or parcel-based) versus continuous (address range-based street segments) reference datasets on achievable match rates has been reported by Zandbergen (2008).

### 3.2.2 Address Validation Best Practices

The address cleaning best practices specifically associated with address validation as described above can be summarized as follows:

- ❖ Address validation should be used during both address standardization and feature matching and interpolation, if a trusted, complete address dataset is available.
- ❖ The temporal footprint of the address validation source should cover the period for which the address in question was supposed to have existed in the dataset. If an assessor parcel database is available, this should be used as an address validation reference dataset.
- ❖ If an address is found to be invalid during address standardization, it should be corrected. If an invalid address is not correctable, it should be associated with the closest valid address.
- ❖ If an address is corrected or assigned to the closest valid address, the action taken should be recorded in the metadata, and the original address should be kept as well.

## 3.3 Address Normalization

Address normalization is defined as the process of identifying the component parts of an address such that they may be transformed into a desired format. This first step is crucial to the input data cleaning process. It is impossible to transform addresses into standard formats or use them for feature matching unless each piece of text is mapped (matched) to its corresponding address attribute. The typical components of a city-style postal address are listed in Table 11. The normalization algorithm must attempt to identify the most likely address attribute to associate with each component of the input address. Many techniques can be applied to this problem. Various approaches ranging from the simplistic to highly advanced in terms of their sophistication will now be described.

### 3.3.1 Substitution-Based Normalization

Substitution-based normalization is the easiest method to implement, as it utilizes lookup tables for identifying commonly encountered terms based on their string values. This simplicity also limits it applicability since by definition it is only capable of substituting correct abbreviations and eliminating (some) extraneous data. This technique makes use of "tokenization", which is defined as the conversion of the string representing the whole address into a series of separate "tokens". The address string is processed from left to right, with embedded spaces separating tokens. One weakness of this method is that the original order of input attributes is extremely critical because of this linear sequential processing. For example, a typical substitution-based normalization process will attempt to populate an internal representation of the parts of the street address listed in Table 11, in that exact order. This method incorporates a set of matching rules which define the valid content each attribute can accept. The matching rules are used in conjunction with lookup tables that list synonyms for identifying common attribute values.

The process of substitution-based normalization proceeds as follows. As each token is encountered, there is an attempt to match it to the next empty attribute in its internal representation, in a sequential order. The lookup tables attempt to identify known token values from common abbreviations such as directionals (e.g. "n" being equal to "North", with either being valid), and the matching rules limit the types of values that can be assigned to each attribute. Unfortunately this simplistic method can easily fail. For instance, when keywords valid for one attribute such as "Circle" and "Drive" are used for others, as in "456 Circle Drive West", with neither in the expected position of a street suffix type.

**Table 11 Common attributes of a city-style postal address (Goldberg 2008)**

| Attribute Components |
|---|
| Number |
| Prefix Directional |
| Street Name |
| Suffix Directional |
| Street Type |
| Unit Type |
| Unit Number |
| Postal Name (Post Office name, USPS default or acceptable name for given USPS ZIP code) |
| USPS ZIP code |
| State |

### 3.3.2 Context-Based Normalization

Context-based normalization is a less commonly applied method that makes use of syntactic (structured rules of a language) and lexical (common usage, meaning of a term) analysis to identify the components of the input address. The strength of this method is its ability to reorder input attributes. In turn, this advantage makes it more complicated and thus harder to implement. This method employs similar steps to those taken by a computer programming language compiler, which is a tool used by programmers to produce an executable file from plain text source code written in a high-level programming language (e.g. C#, Java).

The first step in context-based normalization is referred to as "scrubbing", a process which removes illegal characters and white space from the address input data. The data input string is scanned left to right and all invalid characters are removed or replaced. Punctuation marks, e.g. periods and commas, are all removed, and all white-space characters are collapsed into a single space. All characters are then converted into a single common case, either upper or lower. The next step comprises the lexical analysis**,** where tokenization is performed to convert the scrubbed string into a series of tokens using single spaces as the separator. The order of the tokens remains the same as the original input address. These tokens are then designated as a particular type based on their character content such as numeric, e.g. "1840", alphabetic, e.g. "Century Park", and alphanumeric, e.g. "NO1200". The syntactic analysis is the final step where the tokens are positioned into a parse tree based on a grammar (Table 12). A "parse tree" is a data structure representing the decomposition of an input string into its component parts. The "grammar" is the organized set of rules that describe the language, in this case possible valid combinations of tokens that can legitimately make up an address. These are usually written in Backus-Naur form (BNF), a computer readable notation for describing grammars as combinations of valid components (Table 12).

### Table 12 Backus-Naur form of a postal address (Goldberg 2008)

| Example |
|---|
| <postal-address> ::= <street-address-part> <locality-part> |
| <street-address-part> ::= <house-number> <street-name-part> {"," <suite-number> <suite-type>} |
| <street-name-part> ::= {<pre-directional>} <street-name> <street-type> {<post-directional>} |
| <locality-part> ::= <town-name> "," <state-code> <USPS-ZIP-Code> {"+" <ZIP-extension>} |

The major difficultly in context-based normalization is that the tokens are only typed to the level of the characters they contain, not to the address attributes (e.g. street name, post-directional). However, the address attribute level of token typing can be achieved using lookup tables of common substitutions that can allow tokens to be mapped to address components based on both values and character types. Since it is possible for a single token to be mapped to more than one address attribute, tokens can be rearranged and placed in multiple orders that all satisfy the specified grammar rules. Therefore, constraints must be imposed on the tokens to limit the erroneous assignments. An example would use an iterative method to enforce the original order of the tokens as a first try, then relax the constraint by allowing only tokens of specific types (e.g. numeric, alphanumeric) to be moved in a specific manner. Furthermore, certain keywords can be suppressed in order to minimize their importance. Writing such relaxation rules properly, in the correct order, is the most difficult part of context-based normalization.

### 3.3.3 Probability-Based Normalization

Probability-based normalization is defined as the use of statistical methods to identify the components of the input address. This method is an example of record linkage, the task of identifying features in two or more datasets which essentially refer back to the same feature. Probability-based methods do extremely well at handling difficult cases which require combinations of substitutions, reordering, and removal of extraneous data. A byproduct of being so powerful is the fact that characteristically they are not very easy to implement.

Probability-based normalization algorithms treat address input data as unstructured text that must be semantically annotated to correspond to the appropriate attributes from the target domain, e.g. address attributes. The critical first step in this method is to develop a reference set of candidate features that may possibly match an input feature. This term should not be confused with reference datasets containing the reference features, even though the reference set will most likely be built from them. The reference set defines the search space of possible matches a feature matching algorithm processes to determine an appropriate match. Performing this search becomes more complicated (e.g. data processing time) as the size of the reference set increases. Blocking schemes, which are strategies designed to narrow the set of candidate values (O'Reagan and Saalfeld 1987, Jaro 1989), can be used to limit the size of the search space. After creating a reference set, matches and non-matches between input address elements and their normalized attribute counterparts can be determined. The input elements are scored against the reference set individually as well as collectively using several measures. These scores are combined into vectors and their likelihood as matches or non-matches is determined using such tools as support vector machines, which have been trained on a representative data set (e.g. Michelson and Knoblock 2005).

### 3.3.4 Address Normalization Best Practices

The address cleaning best practices associated with the various methods of address normalization described above can be summarized as follows:

- ❖ Substitution-based normalization should be used as a first step in the normalization process, especially if no other more advanced methods are available.
- ❖ Any deterministic set of rules that creates reproducible results that are certifiably valid should be considered acceptable in substitution-based normalization.
- ❖ In regards to choosing a type or method of normalization, any reproducible technique (e.g. Tokenization, Abbreviation) that produces certifiably valid results should be considered a valid normalization practice.
- ❖ At a minimum, USPS *Publication 28* synonyms should be supported in substitution-based normalization (US Postal Service 2008d).
- ❖ At a minimum, whitespace should be used as a token separator in substitution-based normalization.
- ❖ At a minimum, an exact character-level match should be considered a match in a substitution-based normalization
- ❖ If software can be acquired or developed, context-based normalization should be used.
- ❖ In context-based normalization all alphanumeric characters should be considered valid and exempt from scrubbing.
- ❖ Non-valid (scrubbed) characters should be removed and not replaced with any character when performing context-based normalization.
- ❖ In context-based normalization, any grammar can be used which is based on existing addressing standards, e.g. OASIS xAL Standard (Organization for the Advancement of Struc-

tured Information Standards 2008) or the proposed URISA/FGDC address standard (FGCD 2008b). The grammar chosen should be representative of the address data types the geocoding process is likely to see.

- Only exact case-insensitive character-level token matching should be considered a match in context-based normalization.
- In context-based normalization tokens should be allowed to move no more than two positions from their original location.
- Probability-based normalization should be used if the output certainty of the resulting geocodes meets an acceptable threshold. Experiments should be run to determine what an appropriate threshold should be for a particular cancer researcher. These experiments should contrast the probability of getting a false positive versus the repercussions of such an outcome.
- The score that should be considered a valid match when using probability-based normalization will depend on the confidence which is required by the consumers of the geocoded data. At a minimum, a composite score of 95% or above should be considered a valid match.
- In performing address normalization, at a minimum, researchers and/or staff should be trained to understand as well as perform normalization procedures.

## 3.4 Address Standardization Process

Address standardization is defined herein as the conversion of an address from one normalized format into another. This step is critical, since an accurate, standardized output is the most desirable input to the actual process of geocoding. Address standardization is closely tied to normalization and is greatly affected by the level of performance of any normalization process. Standardization is the process that converts normalized data into a known format required by the various components that comprise the geocoding process. Address standards may be used for different purposes and are likely to vary across organizations since there is no single, prescribed, cross-disciplinary format in use today for constructing address datasets. This variability in formats is a significant barrier to data sharing within the cancer research community. If interoperability is desired, there must be an agreement to implement a standardized format. Table 13 presents two possible address standards by the organization that proposed and/or supports them.

### Table 13 Existing and proposed address standards (Goldberg 2008)

| Organization | Standard |
|---|---|
| USPS | *Publication 28* (US Postal Service 2008d) |
| URISA/ FGDC | *Street Address Data Standard* (United States Federal Geographic Data Committee 2008b) |

To further complicate matters, more than one address standard may be required or in use in a given cancer registry for purposes other than geocoding. Accordingly, after attribute identification and normalization, conversion between common address standards may be required. The most difficult step is writing mapping functions, which are the algorithms that translate between a normalized form and a target output standard. These algorithms transform attributes into the desired formats. Mapping functions implement such tasks as address abbreviation substitution, reduction or expansion, and address attribute reordering, merging, or splitting. The transformations for each attribute in the normalized form must be encoded within the mapping functions. Mapping functions must be defined for each potential standard a geocoder may have to translate an input address into. During feature matching, the input address must be in the same standard as that used for the reference dataset. Consequently, the address standard used by every reference dataset in a geocoder must be supported. This means that a mapping function is required for each address standard. The mapping functions should be defined before the standardization process is initiated, then the appropriate transformations can be executed on the normalized input address. The result is a properly standardized address ready for the feature matching step against the reference data source.

In addition to these technical requirements for supporting address standardization, it is important to note that each cancer registry selects an address standard for their staff to report and record the data (Table 13). For example, NAACCR recommends that when choosing an address standard, registries abide by the data standards in *Standards for Cancer Registries: Data Standards and Data Dictionary* (Hofferkamp and Havener 2008), which references USPS *Publication 28* (US Postal Service 2008d).

## 3.5 Address Standardization Best Practices

The address standardization best practices associated with the aforementioned types of input data and input data standards can be summarized as follows:

- At a minimum, NAACCR standard address data should be able to be geocoded. Ideally, any type of descriptive locational data, both relative and absolute, in any address standard should be an acceptable type of input. This includes any form of postal address, intersections, named places, and relative locations.

- At a minimum, relative input data such as postal street addresses can and should be considered geocodable, as well as relative directional descriptions.
- At a minimum, absolute input data such as E-911 locations (if they are absolute) can and should be considered geocodable, if the appropriate reference dataset is available
- Any reproducible type of standardization technique that produces certifiably valid results should be considered a valid standardization practice.
- At a minimum, all postal address standards for all countries for which geocoding are to be performed should be supported.
- Mapping functions for all supported address standards should be created or obtained.

# 4 Feature Matching

For the purposes of geocoding, there are many different feature matching algorithms available today, each with their own benefits and drawbacks. In general, a feature matching algorithm can be defined as a program that selects the correct reference feature in the reference dataset that represents the input data (Goldberg 2008). A feature interpolation algorithm then uses the selected feature to produce the spatial output. This high level conceptualization is illustrated in Figure 6. Matching algorithms may be non-interactive matching algorithms, for instance they might be automated such that the user is not directly involved. Or, they could be interactive matching algorithms which actively involve the user in making choices when the algorithm fails to produce an exact match. In this scenario the user would either correct or refine the input data, or make a subjective informed decision between two equally likely options.

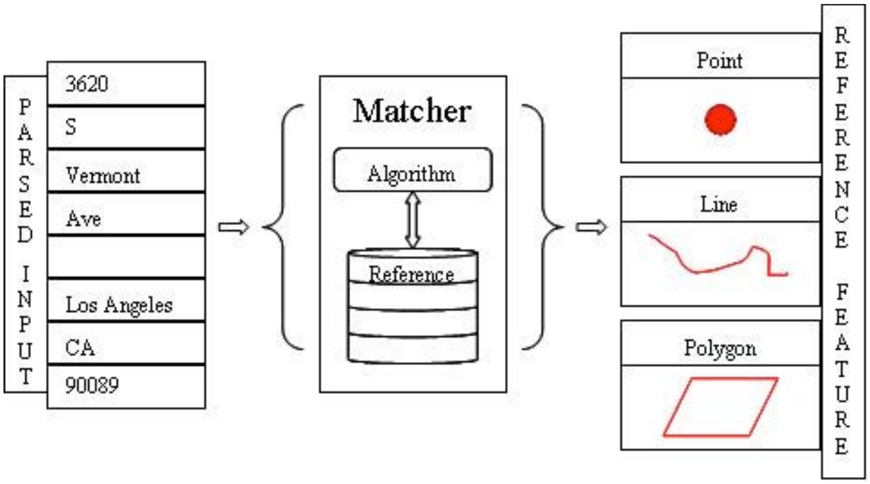


**Figure 6 Generalized feature matching algorithm (Goldberg 2008)**

Most feature matching algorithms operate by constructing and delivering queries defined using Structured Query Language (SQL). The selection attributes are the data that should be returned from the reference dataset in response to a query. These usually consist of the identifiable attributes of the feature such as postal address components, the spatial geometry of the reference feature such as an actual polygon or line segment, and any other preferred descriptive information such as road width or geographic resolution. Data sources include the relational table(s) within the reference dataset that is to be searched. The real power of the query lies in the attribute constraints. Attribute constraints are comprised of zero, one, or more predicates. A predicate is defined as an attribute or value pair defining what the value of an attribute must be for a feature to be selected. Multiple predicates can be linked together with “AND” and “OR” statements to form conjunctions and disjunctions. Case sensitivity is also imperative, because it determines whether or not a database distinguishes between alphabetic characters when evaluating a query against reference features. It is important to note that if case sensitivity is enforced, it can lead to false negatives.

There are two main categories of feature matching algorithms: deterministic and probabilistic (Boscoe 2008). Deterministic matching methods are defined as those based on a series of rules that are processed in a specific progression. These can be thought of as binary operations; a feature is either matched or it is not. Probabilistic matching methods utilize a computational scheme to determine the likelihood, or probability, that a feature matches and returns this value for each feature in the reference set. The normalization processes previously described can be grouped into this same pair of categories. Substitution-based normalization is deterministic, while context- and probability-based are probabilistic. Address normalization is considered to be a higher-resolution version of the feature matching algorithm. While feature matching maps an entire set of input attributes from the input data to a reference feature, address normalization matches each component of the input address to its corresponding address attribute. Both of these processes involve linking records to a reference set. In the case of feature matching the actual features are involved, whereas in the case of normalization the address attributes are utilized. Word stemming and phonetic algorithms are additional methods for encoding the essence of a word that may also be utilized. Word stemming is a technique that reduces a word to its root (stem) and then uses it for essence-level equivalence testing. Phonetic algorithms enable essence-level equivalence testing by representing a word in terms of how it sounds when it is pronounced, e.g. phonetically. For a more detailed discussion on feature matching techniques, see Boscoe (2008) and Goldberg (2008).

It is important to mention that in geocoding practice, less restrictive rules often need to be created and applied in order to match features. This process is commonly referred to as attribute relaxation, defined herein as easing the requirement that all street address attributes must exactly match a feature in the reference data source to obtain a matching street feature. It is usually used in deterministic feature matching since probabilistic methods can account for attribute discrepancies through other processes. Relaxation is generally carried out by removing or altering street address attributes in an iterative manner using a predefined order, thereby increasing the probability of finding a match, though while simultaneously increasing the probability of error.

Lastly, feature matching algorithms also require strings of character data to be compared in order to determine matches and non-matches. There are several methods for computing string comparisons, including character equivalence and essence-level equivalence. Character-level equivalence requires that each character of two strings must be exactly the same. Essence-level equivalence determines whether or not two strings are "essentially" the same. The essence-level technique permits minor misspellings in the input address and returns reference features that "closely match" what the input may have "intended". These methods are applicable to both deterministic and probabilistic feature matching algorithms. Careful attention must always be paid to the accuracy of the results returned when either of these techniques is applied.

## 4.1 Feature Matching Best Practices

The feature matching best practices associated with the use of deterministic and probabilistic feature matching and string matching algorithms can be summarized as follows:

- In performing feature matching, at a minimum, researchers and/or staff should be trained to understand how to create and work with simple database applications.
- Case-sensitivity should not be enforced in feature matching. All data should be converted to upper case as per common data standards, e.g. Hofferkamp and Havener 2008.
- Deterministic matching should be the first feature matching type attempted.
- Any deterministic set of feature matching rules can be used, but they should always be applied in the same order.
- Feature matching rules should be applied in order of decreasing restrictiveness, starting from the most restrictive such that tightly restrictive rules are applied first, and progressively less restrictive rules are applied subsequently upon a previous rule's failure.
- Attribute relaxation should be allowed when using deterministic feature matching.
- Attribute relaxation can be applied in the series of steps and passes, such as those listed in Goldberg (2008).
- Probabilistic matching should be used when deterministic feature matching fails, and if the consumers of the data are comfortable with the confidence threshold.

- When using probabilistic matching, a 95% confidence threshold (at a minimum) should be acceptable.
- Metadata should describe the match probability, when probabilistic feature matching is performed.
- Alternative string comparison algorithms should be used when no exact feature matches can be identified. A two-step approach should be used to compare the original input with the essence-level equivalence match to determine the match and unmatched probabilities (as in the probability-based feature matching approach).
- Both character- and essence-level string comparisons should be supported.
- Character-level equivalence should always be attempted first on every attribute.
- Essence-level equivalence should only be attempted if character-level equivalence fails, and should only be attempted on attributes other than the street name. Only one essence-level equivalence algorithm should be applied at a time. Multiple algorithms can be tried in succession but one should not process the output of the other, e.g. they should both start with the raw data. Metadata should describe the calculated essence of the string used for comparison, and strings that it was matched to in the reference dataset.
- Both stemming and phonetic algorithms should be supported by the geocoding process.
- At a minimum, the Porter Stemmer (Porter 1980) should be supported by the geocoding process.
- At a minimum, the SOUNDEX algorithm should be supported by the geocoding process (see Porter 1980).

## 4.2 Trouble-Shooting Solutions for Dealing with Non-Matches

Feature matching failures can be attributed to two basic causes: ambiguity when matching multiple features and not matching any features (Goldberg 2008). When a failure occurs the address can either remain non-matched and be excluded from a study, or an attempt can be made to reprocess it in some different form or using another method. Excluding non-matchable addresses from a dataset or cancer research project is not recommended (Gregorio et al. 1999, Kwok and Yankaskas 2001, Durr and Froggatt 2002, Bonner et al. 2003, Oliver et al. 2005). These recent studies indicate that significant bias can be introduced if a non-matchable address and the health-related information it represents are excluded. Thus both cancer researchers and cancer registries are advised to re-attempt feature matching through the use of one or more of the following methods: hierarchical geocoding, attribute imputation, composite feature geocoding, or manual review. Hierarchical geocoding is defined herein as using the lower resolution portion of an input address for geocoding. The term feature disambiguation is understood to refer to the attempt to remove ambiguities between ambiguous matches. Attribute imputation has been specified as trying to assign missing data that caused the ambiguity. Composite feature geocoding is defined as obtaining and implementing new reference features based on the ambiguous matches. Manual review occurs when a cancer researcher or registry staff member personally reviews and corrects a non-match.

Hierarchical geocoding is the most frequently used approach to trouble-shooting non-matches. Depending on the reason why geocoding failed in the first place as well as the desired level of accuracy and confidence desired for a particular research study, the lower resolution attribute can be selected. To improve the accuracy associated with choosing lower resolution features, information about the ambiguous features themselves could be utilized. If two or more features returned from the feature matching algorithm have the same level of geographic resolution, the best option is to return the level of geographic resolution which they both have in common. For example, if a the feature matching algorithm returns two streets in the same USPS ZIP code, then a geocode for that USPS ZIP code should be returned. If the two streets are in separate USPS ZIP codes, yet the city is the same, the geocode for the city should be returned. For more information on the implied accuracies within feature hierarchies, see Goldberg (2008).

In feature disambiguation, an attempt is made to determine the correct or best choice out of various possible matches. How this is accomplished depends on why the ambiguity occurred as well as any other information that may be available to help in the choice of the correct one. These cases of ambiguity can result from an error in the reference dataset, or from the input data not being described with enough detail, for instance in a case where a directional field or house number was omitted. Disambiguation is similar to interactive geocoding, thus usually requires the time and subjectivity of a cancer registry staff member. The staff member would choose one of the ambiguous matches as correct, based on other information associated with the input data or by reasoning what they have in common.

Attribute imputation represents another possible approach where missing input address attributes required for geocoding are assigned (Boscoe 2008, Zimmerman 2008). Usually, imputing values will introduce some uncertainty into the resulting spatial output. There is currently no consensus in the literature as to why, how, and under what circumstances imputation should be attempted. At present, whether or not to use this approach is purely a judgment call e.g. by the cancer registry or researcher. The confidence or validity of imputed attributes can increase if the imputed data have been verified from multiple sources. Nevertheless cancer researchers must be cognizant of the greatest possible source of uncertainty that should be associated with spatial output obtained using imputed attributes. And, as the number of imputed attributes increases, the likelihood of error propagation likewise increases. Therefore, imputed values must be identified in the metadata associated with a geocode so a researcher can choose whether or not to utilize a geocode based upon them.

Composite feature geocoding is the next option, if disambiguation through attribute imputation or the subjectivity of a staff member fails. This approach can involve creating a new feature based on ambiguous matches, and using the new feature for interpolation. An example of this approach is the task of delimitating boundaries for imprecise regions (e.g. Reinbacher et al. 2008). Another example is generating a geocode with the quality “midpoint of street segment”. The geocoder basically does the same task, which is to derive a centroid for the bounding box of the conjunction of all ambiguous features. In this case “all ambiguous features” consists of only a single street, and the centroid is derived using a more advanced calculation. However, it may not be possible to obtain the center point of multiple features if the input data do not map to ambiguous features that are topologically connected.

In the future recommendations in the form specific criteria which define when cancer data have met minimum quality standards for geocoding could be established. Such quality control standards would be designed to assist registry staff in making decisions as to when it is appropriate to use specific input datasets for geocoding. Variations of all of the aforementioned trouble-shooting approaches for handling non-matches may or may not be a cost-effective use of cancer researcher or cancer registry resources. If the above approaches to trouble-shooting non-matches all fail, the next option may be to simply hold off on geocoding for the period of time required for the desired reference data sources to be updated, then try running the process again. This option is viable if the researcher or registry staff member believes that the address data are indeed correct, but the reference files are temporally inaccurate, e.g. contain errors and omissions. One example of where this is often the case is in rapidly expanding areas of the country where new construction is being erected. Another case in point occurs in areas where significant reorganization of parcels or streets has occurred, and as a result street names and parcel boundaries have changed since the latest footprint of the reference data was obtained. It also is important to consider using reference data that is most representative of the time period when the addresses were collected. It is of primary importance that the remainder of the input data does not lose accuracy due to the addition of newly updated reference datasets. Nonetheless keeping a record in a non-matched state means that it cannot be included in research or analyses, and should be avoided as often as possible.

The manual review process may be the most accurate as well as the most time consuming method for handling non-matchable addresses. A manual review of a single non-match can take anywhere from a few seconds to a few hours, depending on what exactly caused the mismatch. Examples of quickly solved corrections would be when one of the components of the address attributes is obviously wrong because of incorrect data entry,  such misspelling or address attribute transposition (attributes are in the wrong order). Though such errors might be difficult for a computer program to handle, new advances are being made (Churches et al. 2002, Schumacher 2007, Goldberg et al. 2008a). There are other solutions which require some research but do not involve re-contacting the patient. The task requires querying both the individual address components, combinations, and aliases against different sources, e.g. USPS ZIP+4 database (US Postal Service 2008a), other reference sets, local datasets, address points, and/or parcels. The goal is to either directly identify an error (alias) in the input data, the address or address range in the reference data, or in the patient's name. If the resulting geocode is of such low resolution as to be unusable, the reviewer can manually select the best match at the most reasonable level of geographic resolution. Another option would be to re-contact the input data source (e.g. hospital) if the record is one which requires annual follow-up, since a corrected or updated version may have already been obtained. Lastly, a patient could be contacted in order to obtain corrected or more detailed address information. This would normally only be conducted by individual cancer researchers involved in special studies.

Lastly, it should be pointed out that even though an address may not contain enough or correct content to be directly useful, other information associated with the record may provide information for deducing a more accurate address. For example, if a patient-reported address is not useful, a researcher or registry staff member might be able to connect with a state agency such as the state Department of Motor Vehicles (DMV) or Medicare to acquire a valid address. Connecting to such large, administrative databases can be extremely helpful for enhancing demographic information (e.g. addresses). Nevertheless, large agency-specific databases are designed for administrative purposes and are not built with the intention of enabling surveillance or research. Thus there are limitations to such databases from the viewpoint of cancer researchers which need to be understood. Cancer researchers must be aware of the data collection methods of such sources in order to make correct assumptions when attempting to supplement cancer registry data. Data collection schemes may be different for each cancer registry and will definitely be the subject of state and/or local laws and/or rules. Additional sources of alternative data include phone books and other online sources, which can be accessed for free. Other data sources may require agreements to be made between the registry and private or public institutions. The most common approach is to look for a patient's name in parcel ownership records and associate the address if the match seems reasonable, e.g. if a one-to-one match is found between name and parcel during the time period when the person was known to be living in that city.

## 4.3 Trouble-Shooting Best Practices

Best practices related to trouble-shooting feature matching failures can be summarized as follows:

- All non-matchable addresses should be re-attempted using any of the following approaches: attempt to obtain more information from source; hierarchical geocoding; feature disambiguation; attribute imputation; composite feature geocoding; waiting, e.g. for datasets to be updated; and/or manual review.
- If the geocoding is performed per-record, an unmatched address should be investigated to determine a corrective action after it is processed. If the geocoding process is performed in batch mode, all unmatched addresses should be grouped by type of failure and processed together after the initial processing has been completed.
- In general, the same geocoding process used for the original geocoding attempt should be applied again after the unmatched address has been corrected.
- Any time an ambiguous feature match occurs, only a single feature (which may be a composite feature) should be used for calculating the resulting geocode. If extra information is available which can be used to determine the correct feature, then it should be, and the metadata should record what was used and why that feature was chosen. If additional information is not available and/or the correct feature cannot be identified, a geocode resulting from the interpolation of the lower resolution feature, composite feature, or bounding box should be returned.
- Whether or not to impute missing attribute information will depend on the subjectivity of the cancer researcher or cancer registry. Metadata should be created which indicates which attributes are imputed, the sources used for imputing them, and the original values of any attributes that have been changed.
- Geocoding should be re-attempted at a later date after the reference datasets have been updated when it is obvious that the geocoding failed because the reference datasets were out-of-date (e.g. geocoding an address in a new development that is not present in current versions of a dataset).
- If the time and money are available, manual review of unmatched addresses should be attempted for any and all addresses that are not capable of being processed using automated means.
- If an error (non-match) is obviously a data entry error and the correction is also obvious, it should be corrected by manual or automated (if possible) review and the change noted in the metadata.
- If the geographic resolution of the output geocode is too low to be useful (e.g. county centroid), a person should attempt to reason what better, higher resolution geocode could be assigned based on other information about the patient/tumor (e.g. use city centroid of the diagnosing facility if it is known they visited a facility in their city). Some additional guidance or rules regarding geographic resolution of National Program of Cancer Registries (NPCR) data output geocodes need to be provided.
- If the problem with the input address is not trivially correctable through manual review, alternative sources of information should be reviewed to attempt address correction. If a link-

age to an alternative data source can be determined with a suitable level of certainty, it should be made as long as privacy and confidentiality concerns of the data source are adhered to. Metadata should include the source of the supplemental data, the researcher or cancer registry staff member who made the linkage, the method of linkage (e.g. automatic/manual), the linkage criteria, and the data the linkage was made.
- At a minimum, researchers and/or staff should be trained to understand how to trouble-shoot feature matching failures.

# 5 Conclusions

This research report is the first in a series of three reports on best geocoding practices and provides details on national and local reference datasets that can be used in geocoding, input data standards and rules, including address styles, cleaning normalization and standardization, and various aspects of handling feature matching.

In geocoding practice, there are many different decisions that need to be made, from the beginning to the end of any geocoding processes. These choices usually begin with selecting reference datasets appropriate to the task at hand, and processing (e.g. cleaning) address input data. The tasks progress to deciding how best to validate, normalize and standardize input data. Even more options present themselves when it comes to deciding how best to perform and trouble-shoot feature matching. Therefore the main goal of this report was to encapsulate discussions and concisely assemble the best practices related to these various processes.

Different approaches to the various steps involved in the geocoding process will work best depending on the specific goals of a given research project. When planning or initiating geocoding work, it is important to consider any existing or potential constraints on a study, such as resource limitations, time and budget. It is also critical to allocate sufficient time and resources for production of metadata and training of staff, as it is critical for every step of the geocoding process to be thoroughly documented.

Future work will include clarification of cancer-related data needs, and an evaluation of existing commercial off-the-shelf geocoding systems. These forthcoming efforts will include evaluation and documentation on how easy or difficult different geocoding systems are to work with, for instance with respect to reference data usage, input data handling, and implementation of the various geocoding processes covered in this report.

# 6 Acknowledgements

The development of this research report was funded by the CDC. It draws upon the work of Goldberg (2008), which has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, the Centers for Disease Control and Prevention, and the North American Association of Central Cancer Registries. Daniel Goldberg was also supported by a United States Department of Defense Science, Mathematics, and Research for Transformation Defense Scholarship for Service Program fellowship during portions of the period when this and the earlier report (Goldberg 2008) were prepared.

# 7 References

Abe T and Stinchcomb D 2008 Geocoding Best Practices in Cancer Registries. In Rushton G Armstrong MP Gittler J Greene BR Pavlik CE West MM Zimmerman DL, (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 111-126

ADL 2008 Alexandria Digital Library Gazetteer. WWW Document available at: http://alexandria.ucsb.edu/clients/gazetteer (July 10th 2008)

Beyer KMM, Schultz AF, and Rushton G 2008 Using ZIP Codes as Geocodes in Cancer Research. In Rushton G Armstrong MP Gittler J Greene BR Pavlik CE West MM Zimmerman DL, (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 37-68

Bonner MR, Han D, Nie J, Rogerson P, Vena JE, and Freudenheim JL 2003 Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology* 14(4): 408-411

Boscoe FP 2008 The Science and Art of Geocoding: Tips for Improving Match Rates and Handling Unmatched Cases in Analysis. In Rushton G Armstrong MP Gittler J Greene BR Pavlik CE West MM Zimmerman DL, (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 95-110

Churches T, Christen P, Lim K, and Zhu JX 2002 Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models. *Medical Informatics and Decision Making* 2(9)

Clarke KC 2004 *Getting Started with Geographic Information Systems* 4th Ed., Prentice Hall

County of Sonoma 2008 Vector Data - GIS Data Portal - County of Sonoma. WWW document, https://gis.sonoma-county.org/catalog.asp (July 10th 2008)

Davis Jr. CA and Fonseca FT 2007 Assessing the Certainty of Locations Produced by an Address Geocoding System. *GeoInformatica* 11(1): 103-129

Davis Jr. CA, Fonseca FT, and De Vasconcelos Borges, KA 2003 A Flexible Addressing System for Approximate Geocoding. In *Proceedings of the Fifth Brazilian Symposium on GeoInformatics (GeoInfo 2003)*, Campos do Jordão, São Paulo, Brazil

Durr PA and Froggatt AEA 2002 How Best to Georeference Farms? A Case Study From Cornwall, England. *Preventive Veterinary Medicine* 56: 51-62

ESRI, 2008, Defining the Address Locator Components. ArcGIS 9.2 Desktop Help. WWW document, http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Defining_the_address_locator_components (July 10th 2008)

Goldberg DW 2008 A Geocodes Best Practices Guide. In preparation. Springfield, Il. North American Association of Cancer Registries

Goldberg DW, Wilson JP, Knoblock CA, and Cockburn MG 2008 The Cost of Correcting Geocodes with Manual Resolution. In preparation

Google 2008 Google Earth. WWW document, http://earth.google.com (July 10th 2008)

Grand Valley Metropolitan Council 2008 REGIS: Purchase Digital Data. WWW document, http://www.gvmc-regis.org/data/ordering.html (July 10th 2008)

Gregorio DI, Cromley E, Mrozinski R, and Walsh SJ 1999 Subject Loss in Spatial Analysis of Breast Cancer. *Health & Place* 5(2): 173-177

Grubesic TH and Matisziw TC 2006 On the use of ZIP Codes and ZIP Code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics* 5(58)

Health Level Seven, Inc. 2007 Application Protocol for Electronic Data Exchange in Healthcare Environments, Version 2.6. WWW document, http://www.hl7.org/Library/standards.cfm (July 10th 2008)

Hofferkamp J and Havener L (eds) 2008 *Standards for Cancer Registries: Data Standards and Data Dictionary, Volume II* (12th Edition). Springfield, IL North American Association of Central Cancer Registries.

Hurley SE, Saunders TM, Nivas R, Hertz A, and Reynolds P 2003 Post Office Box Addresses: A Challenge for Geographic Information System-Based Studies. *Epidemiology* 14(4): 386-391

Jaro M 1989 Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 89: 414-420

Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian SV, and Carson R 2002 ZIP Code Caveat: Bias Due to Spatiotemporal Mismatches Between ZIP Codes and US Census-Defined Areas: The Public Health Disparities Geocoding Project. *American Journal of Public Health* 92(7): 1100-1102

Kwok RK and Yankaskas BC 2001 The Use of Census Data for Determining Race and Education as SES Indicators A Validation Study. *Annals of Epidemiology* 11(3): 171-177

Lockyer B 2005 Office of the Attorney General of the State of California Legal Opinion 04-1105. WWW document, http://ag.ca.gov/opinions/pdfs/04-1105.pdf (July 10th 2008)

Los Angeles County Assessor 2008 LA Assessor - Parcel Viewer. WWW document, http://assessormap.co.la.ca.us/mapping/viewer.asp (July 10th 2008)

Michelson M and Knoblock CA 2005 Semantic Annotation of Unstructured and Ungrammatical Text. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland

Microsoft 2008 Microsoft Virtual Earth. WWW document, http://www.microsoft.com/VirtualEarth/ (July 10th 2008)

NAACCR 2007 Data Standards and Data Dictionary. Standards for Cancer Registries, Vol. II. Chapter 10: Data Dictionary. Twelfth Ed., Record Layout Version 11.2 North American Association of Central Cancer Registries, 419 p

NAVTEQ 2008 NAVSTREETS. WWW document, http://developer.navteq.com /site/global/dev_resources/170_navteqproducts/navdataformats/navstreets/p_navstreets.jsp (July 10th 2008)

Oliver MN, Matthews KA, Siadaty M, Hauck FR, and Pickle LW 2005 Geographic Bias Related to Geocoding in Epidemiologic Studies. *International Journal of Health Geographics* 4(29)

O'Reagan RT and Saalfeld A 1987 Geocoding Theory and Practice at the Bureau of the Census. Statistical Research Report Census/SRD/RR-87/29. Washington, DC, United States Bureau of Census

Organization for the Advancement of Structured Information Standards 2008 OASIS xAL Standard v2.0. WWW document, http://www.oasis-open.org/committees/ciq/ download.html (July 10th 2008)

Porter MF 1980 An algorithm for suffix stripping, *Program* 14(3): 130-137

Reinbacher I, Benkert M, van Kreveld M, Mitchell JSB, and Wolff A 2008 Delineating Boundaries for Imprecise Regions. *Algorithmica* 50(3): 386-414

Rushton G, Armstrong, MP, Gittler J, Greene BR, Pavlik CE, West MW, and Zimmerman DL (eds) 2008 *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*, Boca Raton, Fl CRC Press

Schumacher S 2007 Probabilistic Versus Deterministic Data Matching: Making an Accurate Decision. *DM Direct* Special Report (January 18, 2007 Issue). WWW document, http://www.dmreview.com/article_sub.cfm?articleId=1071712 (July 10th 2008)

Shi X 2007 Evaluating the Uncertainty Caused by P.O. Box Addresses in Environmental Health Studies: A restricted Monte Carlo Approach. *International Journal of Geographical Information Science* 21(3): 325-340

Stage D and von Meyer N 2005 An Assessment of Parcel Data in the United States Survey Results. Washington, DC, US Federal Geographic Data Committee Subcommittee on Cadastral Data. WWW document, http://www.nationalcad.org/showdocs.asp?docid=170 (July 10th 2008)

Tele Atlas Inc. 2008a Dynamap Map Database. WWW document, http://www.teleatlas.com/OurProducts/MapData/Dynamap/index.htm (July 10th 2008)

Tele Atlas Inc. 2008b MultiNet Map Database. WWW document, http://www.teleatlas.com/OurProducts/MapData/Multinet/index.htm (July 10th 2008)

US Board on Geographic Names 2008 *Geographic Names Information System*. Reston, VA United States Board on Geographic Names. WWW document, http://geonames.usgs.gov/pls/gnispublic (July 10th 2008)

US Census Bureau 2008a *American Community Survey*, Washington, DC, United States Census Bureau. WWW document, http://www.census.gov/acs. (July 10th 2008))

US Census Bureau 2008b *MAF/TIGER Accuracy Improvement Project*. Washington, DC, United States Census Bureau. WWW document, http://www.census.gov/geo/mod/maftiger.html (July 10th 2008)

US Federal Geographic Data Committee, 2008a *The Federal Geographic Data Committee*. WWW document, http://www.fgdc.gov/ (May 1st 2008)

US Federal Geographic Data Committee 2008b *Content Standard for Digital Geospatial Metadata*. WWW document, http://www.fgdc.gov/metadata/csdgm (July 10th 2008)

US National Geospatial-Intelligence Agency 2008 *NGA GNS Search*. Bethesda, MD United States National Geospatial-Intelligence Agency. WWW document, http://geonames.nga.mil/ggmagaz/geonames4.asp (July 10th 2008)

US Postal Service 2008a *Address Information System Products Technical Guide*. Washington, DC, United States Postal Service. WWW document, http://ribbs.usps.gov/files/Addressing/PUBS/AIS.pdf (July 10th 2008)

US Postal Service 2008b *CASS Mailer's Guide*. Washington, DC United States Postal Service. WWW document, http://ribbs.usps.gov/doc/cmg.html (July 10th 2008)

US Postal Service 2008c *Locatable Address Conversion System*. Washington, DC, United States Postal Service. WWW document, http://www.usps.com/ncsc/addressservices/ addressqualityservices/lacsystem.htm (July 10th 2008)

US Postal Service 2008d *Publication 28 – Postal Addressing Standards*. Washington, DC, United States Postal Service. WWW document, http://pe.usps.com/text/pub28/welcome.htm (July 10th 2008)

US Postal Service 2008e LACSLink Product. USPS Products, Management and Services. WWW document, http://ribbs.usps.gov/psp/psp/AQ/LACSLink.htm (July 10th 2008)

Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, and Hartge P 2005 Positional accuracy of two methods of geocoding. *Epidemiology* 16(4): 542-547

Zandbergen PA 2008 A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32(3): 214-232

Zimmerman DL 2008 Statistical Methods for Incompletely and Incorrectly Geocoded Cancer Data. In Rushton G Armstrong MP Gittler J Greene BR Pavlik CE West MM Zimmerman DL, (eds) *Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. Boca Raton, Fl CRC Press: 165-180

# 8 List of Terms

| Abbreviation | Description |
|---|---|
| ADL | Alexandria Digital Library |
| BGN | United States Board on Geographic Names |
| CDC | Centers for Disease Control and Prevention |
| DCPC | Division of Cancer Prevention and Control |
| DMV | Department of Motor Vehicles |
| E-911 | Emergency 911 |
| FCC | Feature Classification Code |
| FGDC | Federal Geographic Data Committee |
| GPS | Global Positioning System |
| HC | Highway Contract |
| IR | Information Retrieval |
| MCD | Minor Civil Division |
| NAACCR | North American Association of Central Cancer Registries |
| NGA | National Geospatial-Intelligence Agency |
| NGC | Northrop Grumman Corporation |
| NPCR | National Program of Cancer Registries |
| PO Box | Post Office Box |
| RR | Rural Route |
| SQL | Structured Query Language |
| TIGER | Topographically Integrated Geographic Encoding and Referencing |
| URISA | Urban and Regional Information Systems Association |
| US | United States |
| USPS | United States Postal Service |
| ZCTA | ZIP Code Tabulation Area |