Address Standardization

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

GIS research laboratory

Technical Report No. 12

Ecological Consequences of Artificial Night Lighting
A Geocoding Best Practices Guide
The Handbook of Geographic Information Science
Geographic Information Systems and Public Health
Eliminating Perinatal Disparity

Table of Contents

Executive Summary

This technical report outlines the details of contemporary address standardization techniques and the current implementation of address standardization within the USC WebGIS Open Source Geocoding Platform. Different levels or degrees of address standardization are discussed in the context of the address data cleaning process, as well as other procedures which comprise address standardization. Commonly collected locational input data which requires standardization are described, including examples which illustrate other essential data cleaning processes such as parsing and normalization. Address validation and normalization procedures are covered in some depth since these activities are considered critical components of the address cleaning process. The current status of the USC WebGIS Open Source Geocoding Platform is summarized, with particular attention to currently implemented address standardization procedures. Lastly, the best path forward with the intent of improving the current state of address standardization practices is presented, as a set of priorities or recommendations that, if implemented, can ensure high quality in standardized addresses.

1. Introduction

Geocoding is most commonly considered to be the process of converting a locational description such as a street address into some form of geographic representation such as geographic coordinates (latitude and longitude). This process is critical in many scientific arenas as it is typically the first step used to create the spatial data employed in subsequent spatial analyses. Accordingly, the accuracy, granularity, and reliability of geocoded data are of paramount importance in studies that use address data as their underlying geospatial data source. To this end, the USC GIS Research Laboratory has undertaken a multi-year effort to develop a scalable, reliable, accurate and extensible geocoding platform for use in the academic and larger scientific communities. Address standardization, which can be summarized as the conversion of an address from one format into another, is a critical component of geocoding which has been fully integrated in the USC WebGIS Open Source Geocoding Platform. A high-level overview of the relationships between the various components in a geocoding system, including where address standardization fits into such architectures is displayed in Figure 1.

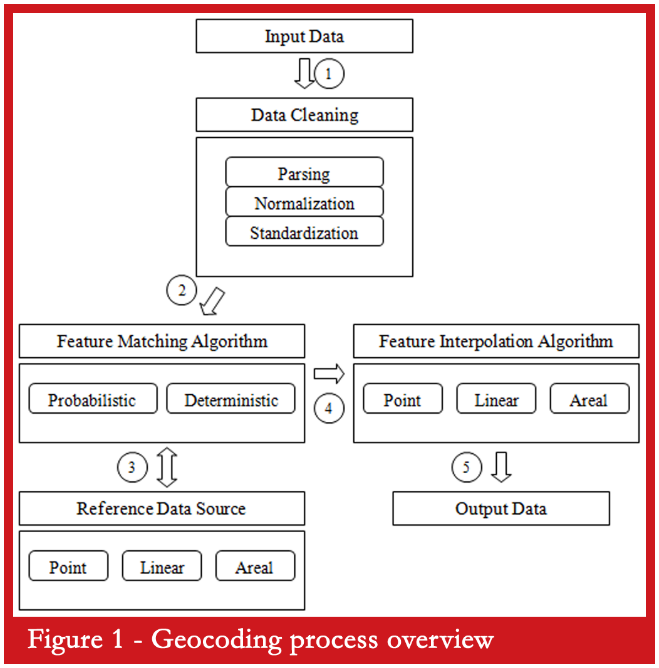

Figure 1 - Geocoding process overview

The purpose of this report is three-fold: (1) to describe the range of potential interactions in geocoding that employ address standardization; (2) to describe the solution that is currently implemented in the USC WebGIS Open Source Geocoding Platform.; and (3) propose an acceptable level of address standardization and how the outcomes are influenced and/or affected by other geocoding decisions and/or processes.

2. Different Levels of Address Standardization

Address standardization is one of several procedures that constitute the data cleaning process (Figure 1). How "clean" input data is may be one of the greatest contributing factors to the success or failure of producing a successful geocode (Goldberg, 2008 [1]). Address data are considered "dirty" for several reasons, such as the use of non-standard abbreviations and attribute orderings, or even simple data entry mistakes.

Address standardization is a crucial step in the geocoding process, as well as in many other fields of interest which require accurate addresses. These other fields cover a vast range of government, industry and scientific arenas of study, including health, natural and man-made hazards, the environment and farming, education, land-use planning, law enforcement, education, etc. The interests of these disciplines overlap when they require the development of significant (i.e. large volume) highly accurate locational data sets to support their efforts, sometimes in the form of large databases. Address standardization comes into play as a fundamental step in compiling such resources since a substantial fraction of the address information that is collected is incomplete and/or ambiguous.

In addition to address standardization, the input data cleaning steps also include address normalization and parsing. Address normalization identifies the component parts of an address, while parsing is usually treated as the part of the normalization algorithm that attempts to identify the most likely address attribute to associate with each component of the input address. The main goal of the address parser is to break an unformatted input street address (e.g. "123 No Main Street") into its separate components and format each into its respective standard format according to USPS Publication 28 (e.g. "123", "N", "Main", and "ST") (U.S. Postal Service, 2009 [2]). Over the last several decades a significant amount of Computer Science research has been invested in addressing the challenge presented by parsing. Many computational techniques can be applied to this problem, and examples from the simplistic to highly advanced can be found in Goldberg (2008 [1]). One recent example can be found in Christen et al. (2006 [3]), where a geocoding system was developed that incorporated a learning address parser based on hidden Markov models to segment free-form addresses into components, coupled with a rule-based matching engine to determine the best matches to the reference dataset.

The following subsections in this report describe the range of addresses that may be encountered, and how address standardization can help fill in gaps and resolve ambiguities in the address input data.

2.1 The Input Address Data

Locational input data collected in various disciplines comes in many forms and with varying degrees of completeness. Low quality or incompleteness in input data represents one of the most significant problems faced by those who require accurate address data. Thus a primary aim of address collection systems should be to eliminate the ambiguities that are introduced as the basic elements or attributes of address data are, often accidentally, removed from addresses. The best-possible-case scenario would be to gather the most complete information possible at the very beginning of the process, in other words at the source of any address data collection activity. The desired level of completeness in input address data will be herein referred to as the "gold-standard" in postal address data. The practice of transforming an incompletely described address into a (completely described) gold standard address is performed by most commercial geocoders, as evidenced by the inclusion of the full attributes of the matched feature (address) generally included with the geocode result. Feature matching is where a single feature represents only a single real world entity, e.g. a point feature, as opposed to a feature which represents a range or series of real world entities, e.g. a line feature.

The most common address attribute components encountered in address standardization processes include street name, number and type, prefix and suffix directionals, unit type and number, postal name (Post Office name, USPS default or acceptable name for given USPS ZIP Code), USPS ZIP Code and state. An address which contained all of this information would illustrate what can be considered the "gold-standard" in postal address data (Churches et al. 2002

[4], Goldberg 2008 [1]). The following example of a "gold-standard" address contains valid information in each of the possible attribute fields and indicates enough information to produce a geocode down to the sub-parcel unit or the floor level:

> "3620 ½ South Vermont Avenue East, Unit 444, Los Angeles, CA, 90089-0255"

For instance, in the geographic scale progression used during the feature matching step in a geocoding algorithm, a search for this address is first confined by a state, then by a city, then by a detailed USPS ZIP Code to successively limit the number of possible candidate features to each of these areas. Next street name ambiguity is removed by the prefix and suffix directionals associated with the name, "South" and "East", respectively, as well as the street type indication, "Avenue". Parcel identification is then possible through the use of the street number, "3620", assuming that a parcel reference dataset exists and is accessible to the feature matching algorithm. Next, a three dimensional (3D) geocode can finally be produced from the sub-parcel identification by combining the unit indicators, "½" and "Unit 444" 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 "444" can mean different things in different localities, e.g. they can both refer to subdivided parcels, subdivisions within a parcel, or even lots in a trailer park.

The aforementioned example above illustrates the best-possible-case scenario in terms of postal address specification, but it is rarely encountered. In geocoding practice, reference dataset availability is also critical in producing high quality address data. A reference dataset such as street or parcel-based data is the underlying geographic database containing geographic features that a geocoder can use to generate a geographic output. Unfortunately, high quality reference datasets do not exist for many large regions, and details such as the floor plan within a building are seldom available. Also, input data are hardly ever specified this completely at the original source of data collection. 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 (see next paragraph for a more detailed description) and must be used in conjunction with other sources to obtain specially precise output geocodes, which may still be subject to some error.

Address validation, another important component of the address cleaning process, determines if an input address corresponds to a location that actually exists in the real world. The simplest way to attempt address validation is to perform feature matching using a reference dataset containing discrete features. Address validation is currently impractical, because although a simple approach would be to use a USPS CASS certified product to validate each address, CASS systems are prohibited from validating segment-like reference data because of bulk mailers. Thus parcel or address point reference data must still be used, such as the commonly used USPS ZIP+4 database (United States Postal Service 2008 [5]). Other sources such as assessor parcel files may be available for different areas and may provide additional help. Note that even though some addresses may validate, they still may not be geocodable due to problems or shortcomings with the reference dataset.

While parcel data have proven useful for improving upon address data, it should be noted that in most counties, assessors are under no mandate to include the situs address of a parcel (the actual physical address associated with the parcel) in their databases. In these cases, the mailing address of the owner may be all that is available, and this may or may not record the actual address of the actual parcel. Address validation will become more feasible in the future assuming E911 address points are available, as an alternative and better option for performing address validation.

2.2 Address Normalization

Address normalization can be considered a precursor step to successful address standardization. This step generally consists of identifying the component parts or attributes of an address so that they may be transformed into some other desired format. A normalization algorithm must attempt to identify the most likely address attribute to associate with each component of an input address with respect to the "gold standard", as described above. That said, it is clear for all to see that address normalization is critical to the address cleaning process. Hence, without identifying which piece of text corresponds to which address attribute, it is impossible to subsequently transform them between standard formats or use them for feature matching. A range of normalization approaches may be utilized in geocoding practice, including substitution, context and probability-based normalization. Examples of substitution and context-based normalization are presented herein.

Substitution-based normalization makes use of lookup tables for identifying commonly encountered terms based on their string values. This is the most popular method because it is the easiest to implement. This simplicity also makes it applicable to the fewest number of cases, for instance by substituting correct abbreviations and eliminating (some) extraneous data. In this method, "tokenization" converts the string representing the whole address into a series of separate "tokens" by processing it left to right, with embedded spaces being used to separate tokens. The original order of input address attributes is critical because of this linear sequential processing. A typical geocoding system, for example, will endeavor to populate an internal representation of the parts of the street address described above. A set of matching rules define the valid content each attribute can accept, and are used in conjunction with lookup tables that list synonyms for identifying common attribute values. As each token is encountered, the system tries 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 (i.e. "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. To illustrate how it works, the following address will be processed, matching it to the address attributes listed above:

> "3620 Vermont Ave, RM444, Los Angeles,
> CA 90089"

In the first step, a match is attempted between the first token of the address, "3620" and the internal attribute in the first index, "number". This token satisfies the matching rule for this internal attribute, i.e. that the data must be a number, and it is therefore accepted and assigned to this attribute. Next, a match is attempted between the second word, "Vermont", and the address attribute that comprises the second index, the pre-directional. This time, the match will fail because the matching rule for this attribute is that data must be a valid form of a directional, which this word is not. The current token "Vermont" is then attempted to be matched to the next attribute (index 3, street name). The matching rule for this has no restrictions on content, so the token is assigned. The next token, "Ave", has a match attempted with the valid attributes at index 4 (the post-directional) which fails. Another match is attempted with the next address attribute at the next index (5, street type), which is successful so it is assigned. The remainder of the tokens are subsequently assigned in a similar manner. It is easy to see how this simplistic method can easily get into trouble when keywords valid for one attribute such as "Circle" and "Drive" are used for others as in "123 Circle Drive West", with neither in the expected position of a street suffix type.

Context-based normalization makes use of syntactic and lexical analysis to identify the components of the input address. The main benefit of this less commonly applied method is its support for reordering input attributes. This also makes it more complicated and harder to implement. Context-based normalization consists of steps very similar to those taken by a programming language compiler, a tool used by programmers to produce an executable file from plain text source code written in a high-level programming language.

The first step is called “scrubbing”, which removes illegal characters and white space from the input datum. The input string is scanned left to right and all invalid characters are removed or replaced. Punctuation marks such as 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, referred to as lexical analysis, breaks the scrubbed string into typed tokens. 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 input address. Referring back to the substitution example above, these tokens are then assigned a type based on their character content such as numeric: “3620”, alphabetic: “Vermont”, and alphanumeric: “RM444”. The final step, syntactic analysis, places the tokens into a parse tree based on a grammar. This 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 notation for describing grammars as combinations of valid components. An example of an address described in BNF is as follows:

```
<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>}
```

In this example, a postal address is composed of two components the street-address-part and the locality-part. The street-address-part is composed of a house-number, a street-name-part, and an optional suite-number and suite-type, which would be preceded by a comma if they existed. The remaining components are composed in a similar fashion.

The difficult part of context-based normalization is that the tokens described thus far have only been typed to the level of the characters they contain, not to the domain of address attributes, such as street name. This level of domain-specific token typing can be achieved using lookup tables of common substitutions that map tokens to address components based on both character types and values. It is possible for a single token to be mapped to more than one address attribute. Thus, these tokens can be rearranged and placed in multiple orders that all satisfy the grammar. Therefore constraints must be imposed on them to limit erroneous assignments. Possible options include using an iterative method to enforce the original order of the tokens as a first try, then relaxing the constraint by allowing only tokens of specific types to be moved in a specific manner, etc. Also, the suppression of certain keywords can be employed such that their importance or relevance is minimized. Thus the most difficult part of performing context-based normalization is writing these relaxation rules properly, in the correct order. One must walk a fine line and carefully think about what one should do to which components of the address in what order, otherwise the tokens in the input address might be moved from their original position and seemingly produce “valid” addresses that misrepresent the true address.

Probability-based normalization makes use of statistical methods to identify the components of an input address. It derives mainly from the field of machine learning, a branch of Computer Science dealing with algorithms that induce knowledge from data. In particular, it is an example of record linkage, the task of finding features in two or more datasets which essentially refer to the same feature (Winkler 1995 [6], Jaro 1995 [7], Churches et al. 2002 [4]). Record linkage is utilized to create a frame, remove duplicates from files, or to combine files so that relationships in two or more data elements from disparate sources can be studied. A detailed account of various computer-assisted record linking methods is provided in Winkler (1995 [6]) and Churches et al. (2002 [4]) describe an alternative approach to address standardization, using a combination of lexicon-based tokenization and probabilistic hidden Markov models. Methods such as these excel at handling the difficult cases; those which require combinations of substitutions, reordering, and removal of extraneous data. Being so powerful, they are typically very difficult to implement, and are usually seen only in research scenarios. Probabilistic algorithms essentially treat the input address as unstructured text that needs to be semantically annotated with the appropriate attributes from the target domain, i.e. address attributes. The key to this approach is the development of an optimal set of candidate features that may possibly match an input feature. The optimal set of candidate features defines the search space of possible matches a feature matching algorithm processes to determine an appropriate match. In most cases the complexity of performing this search (i.e. processing time) grows linearly with the size of the reference set. In the worst case, the search space can be composed of the entire optimal set of candidate features, resulting in non-optimal searching. The intelligent use of blocking schemes, or strategies designed to narrow the set of candidate values (O'Reagan and Saalfeld 1987 [8], Jaro 1989 [9]), can 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 (SVMs), which have been trained on a representative data set. For complete details of a practical example using this method see Michelson and Knoblock (2005 [10]).

2.3 Address Standardization

Address standardization can, in the narrowest sense, be defined as the conversion of an address from one normalized format into another. It is closely linked to normalization and is heavily influenced by the performance of the normalization process. In a nutshell, standardization converts the normalized data into the correct format expected by the subsequent components of an address processing system, such as a geocoder. Address standards may be used for different purposes and may vary across organizations since there is no single, set format; unfortunately, this variability in formats presents a barrier to data sharing among organizations. Interoperability assumes an agreement to implement a standardized format. One of the major hurdles to overcome in implementing an address standardization system is that more than one address standard may be required or in use by different entities (government, academia, etc.) for many purposes, including those outside of the geocoding process. Therefore, after attribute identification and normalization, transformation between common address standards may be required. In addition to technical requirements for address standard support, address collection and usage entities (i.e. cancer registries) must select an address standard for their staff to report and record the data in. The existing and proposed address standards include the following:

- TIGER®, TIGER/Line® and TIGER®-Related Products 2008. Topologically Integrated Geographic Encoding and Referencing System (United States Census Bureau 2009 [11])
- USPS - Publication 28 - Postal Addressing Standards. (United States Postal Service 2008 [2])
- Urban and Regional Information Systems Association (URISA)/United States Federal Geographic Data Committee (FGDC) - Street Address Data Standard (United States Federal Geographic Data Committee 2008 [12])

The difficult portion of this process is writing the "mapping functions", which are the algorithms that translate between a normalized form and a target output standard. These functions transform attributes into the desired formats by applying such tasks as abbreviation substitution, reduction or expansion, and attribute reordering, merging, or splitting. These transformations are encoded within the mapping functions for each attribute in the normalized form. Mapping functions must be defined a priori for each of the potential standards the address processing system (i.e. geocoder) may have to translate an input address into, and there are commonly many. To understand this, consider that during feature matching the input address must be in the same standard as that used for the reference dataset before a match can be attempted. Therefore, the address standard used by every reference dataset in a geocoder must be supported, i.e. a mapping function is required for each reference dataset. With the mapping functions defined a priori, the standardization process can simply execute the appropriate transformation on the normalized input address, and a properly standardized address ready for the reference data source will be produced.

In terms of the implications of choosing one standard over another, as an example, the U.S. Census Bureau is currently utilizing the URISA/FGDC standards for all their address standardization activities and has integrated the most recent version of TIGER/Line® data into their geocoding systems. In general, the USPS – Publication 28 standard is considered to be less detailed than the URISA/FGDC standard, such that choosing the former for address standardization activities locks the user into the production of a standardized dataset that may be less comprehensive than if the URISA standard were employed.

3. Status of the USC Geocoding Platform

The current status of the USC WebGIS Open Source Geocoding Platform is described in detail in Goldberg (2009 [13]; Figure 1). The system consists of a series of independent, reusable software components that are implemented online via a graphical web user interface that allows a user to geocode single records as well as databases of records in batch mode. The USC GIS Research Laboratory has also developed a set of web APIs that can be used by a user or user-written programs to send address data to the USC GIS Research Laboratory to be geocoded and returned.

With respect to the level of address standardization currently implemented in this system, all of the input data cleaning components of a traditional geocoding system are implemented, including address normalization and parsing. The system accepts input data supplied by a user in the form of an unparsed street address and a city and/or USPS ZIP code combination. The input address entered by a user usually consists of an unparsed street address, along with a city, state, and USPS ZIP code including the +4 portion of a ZIP+4. The address parsing and normalization component is a non-USPS CASS certified deterministic token-based system that processes tokens left-to-right based on white space separation using synonym tables of common term values and a context aware state machine to determine token type and normalized value. This unparsed input street address is first parsed and normalized to identify standard values for each of the postal address components. Parsing and normalization are applied to the street address portion of an address including the secondary unit and can recognize PO Boxes and other delivery route address types, e.g. Rural Routes. Addresses are standardized to the USPS Publication 28 specification (U.S. Postal Service, 2009 [2]).

As part of the address standardization process, the USC WebGIS Open Source Geocoding Platform implements a complete enumeration of the USPS Publication 28 accepted postal address components and abbreviations, as well as those not in the standard but still commonly used. A synonym matching system is implemented that uses hash tables (for quick access) to identify the possible postal attribute types for each of the words of the input address. The input address is first tokenized on white space and the set of tokens are processed linearly from left to right. As each token is encountered, the possible types are identified using the synonym matcher and the correct one is chosen based on the position in the token set and the attributes that already have been identified. This implementation is wrapped into a standalone component that can be used in isolation or integrated into other software systems.

The USC WebGIS Open Source Geocoder attempts to find one or more reference features that match the input address from within each of the reference data layers that it maintains once the address standardization process is completed. If the system is able to obtain a matching reference feature, feature interpolation is performed to determine an appropriate output location within or along the reference feature based on the input address. The output geocode contains the geographic coordinates (latitude and longitude) of the calculated location as well as metadata that provide information on the selection process and criteria used with the reference feature and the probable quality of the geocoded.

The USC WebGIS Open Source Geocoder has been used by more than 1,600 registered users to geocode over 7,000,000 addresses in all 50 states to date. While actual per-record processing time varies and is entirely dependent on the number of attempts the feature matching algorithms must attempt, i.e. is a match found on the first query in the first reference data source or does the system have to try all versions of all queries (i.e. complete relaxation with both soundex and substring matching) across all reference data sources, the average processing time for a single geocode is 0.3 seconds. The system averages 60,000 geocoding queries per day, with upwards of 10 queries being processed at any one instant in time.

4. Recommendations

The best path forward with the intent of improving the current state of address standardization practices in geocoding systems would be to extend the capabilities of these systems as suggested by the following recommendations. The three innovations with the potential for the most drastic impacts and improvements on the address standardization processes are presented. Each of these recommendations is discussed in terms of the rationale behind it and the tradeoff between the level of benefit that can potentially be realized within the cancer registry community and the potential costs associated with its particular implementation or adoption.

4.1. Adopting a Single Addressing Standard

The first and most crucial step toward improving address standardization processes involves the selection and universal adoption of a single address standard to be utilized consistently across all aspects of registry operations. A consensus needs to be brokered between all parties as to the fundamental schema used to represent postal address data before the true benefits of subsequent address standardization processes can be realized. Therefore, it is recommended that all data collection procedures, registry operations, and registry formatting standards be updated and/or altered to use a consistent postal address standard. The recommended standard is the URISA/FGDC address standard (United States Federal Geographic Data Committee 2008 [12]) for three reasons. First, this is the most comprehensive standard available and is a superset of the USPS Publication 28 standard (United States Postal Service 2008 [2]). The USPS Publication 28 standard is currently the most commonly used, and can be easily derived as a special case of the URISA/FGDC address standard, while the reverse is not as easily obtainable. Second, the most current version of the TIGER/Line® data utilizes an address standard which is compatible with this URISA/FGDC address standard. These files are the most popular geocoding reference data source used, and therefore the address standard chosen should be compatible with them to ensure the widest level of geocoder compatibility. Finally, the URISA/FGDC address standard is recognized and recommended as the address standard of choice by federal agencies and national organizations specializing in the production, transmission, and utilization of many forms of geographic data. Health data in general and cancer incidence data in particular are one of the most critical types of spatial data, capable of revealing non-random spatial distributions that are indicative of serious environmental and/or cultural problems. For cancer registry data to fit seamlessly into the larger context of interoperable government data, useful at a minimum in disease surveillance, the same address standard needs to be universally applied to enable linking across the full spectrum of available data.

The transition to a single address standard will most likely require significant financial costs to alter current standards, retrain personnel in new practices, and re-implement portions of existing software systems. However, without this most basic agreement between data collectors, aggregators, consumers and software providers, it is inevitable that the problems with address standardization and interoperability seen in current systems will continue indefinitely.

4.2 Moving Standardization Processes Closer to Data Originators

The second critical task required to improve address standardization practices is to move them as close as possible to the originator of the data. Currently, the majority of address standardization takes place at centralized registries which are typically far removed from the point where the data was originally collected. This is troublesome for three reasons. First, problem cases such as missing, ambiguous, or incorrect data may not be solvable or correctable beyond the point of initial data collection. These circumstances usually require the local-level knowledge present at the point of collection about the particulars of the addressing system in a region or the ability to ask the person providing the information for immediate clarification to solve them. Without these abilities, such problem instances may never be correctable without a person or software system making an assumption which may or may not be correct. Second, a substantial amount of time may have passed between data collection and address processing potentially resulting in out-of-date information being used to perform the standardization which could have been overcome had the data been standardized immediately upon collection. Finally, the staff member entering the address information into the tumor abstract, for example, may be several organizational steps removed from where the data were originally collected (or even completely separated) with no way to obtain any further information that could be useful in the standardization process.

Because of these limitations inherent to the current logical data flow from data collection in hospitals, clinics, and treatment centers to address standardization occurring at central registries, it is recommended that address standardization, including address normalization and validation, be performed immediately upon data collection. As address data are collected and/or entered into patient records within hospital, clinic and facility databases, the staff member performing the ingest should be immediately notified of incorrect, incomplete, or potentially ambiguous address data. This will allow the staff member to take preventive action to correct the erroneous or suspect address data by utilizing their local-level knowledge of the region or asking the data provider (the patient or their representative) to clarify the problem data.

Changing the organizational level at which address standardization is performed and the way in which address data are collected will present many challenges, both operationally and financially. First, because address standardization is often performed as part of the geocoding process its utilization at the point of data collection may be prohibited because of the higher level of cost associated with the reference data sets or confidentiality/privacy issues inherent in sending data across networks to be processed as happens today with many geocoding procedures. To overcome this hurdle, software providers will need to separate these processes such that address standardization can be performed independently of the geocoding process. Second, data collection procedures will need to be augmented to include an additional address review standardization step. This will require new systems to be put in place to capture address data in digital form when a person initially identifies themselves at a hospital, clinic, or treatment facility (rather than just paper). Finally, staff members will need to be trained in address processing techniques so that they may be able to recognize and remedy problem cases while the person is still available to provide clarification.

4.3 Adopting a Single Open Source Address Standardization System

The third and final step required to make significant strides toward improving address standardization processes across the cancer registry community is the development, dissemination, and adoption of a single address processing system that can be used by all hospitals, clinics, treatment facilities, and registries across the country. Currently, each of the many organizations responsible for the production of the standardized addresses found in cancer registry data typically maintains its own internal processes, procedures, and software systems for this task. If and when address standardization procedures are distributed away from the central registries and moved closer to the originators of the data, the multiplicity of systems and processes (i.e. the status quo) will be much more difficult to sustain. Although many commercial address processing and geocoding systems currently in use may be USPS CASS certified, meaning they have met or exceeded a certain specific set of accuracy criteria, they are typically closed-source solutions with their internal workings hidden from the user, each with its own algorithms, assumptions, and accuracy levels. This means that it is difficult to determine if one address standardization system is comparable or compatible with another, potentially resulting in consolidated records with inconsistent levels of address standardization quality.

To prevent inconsistencies from appearing in consolidated data, it is therefore recommended that all address standardization be performed with the same set of software tools. This recommendation will require that a set of address standardization software tools be developed that could be utilized by many different organizations. For widespread adoption to become a reality, this system would need to be freely available to limit the upfront costs of acquiring the system as organizations will need to expend time and effort to incorporate it into their existing address processing workflow and/or create such workflows for the first time. Furthermore, the system would need to be open source, allowing all potential users to inspect the internal workings of the system to ensure it functions as well as their current approaches and is compatible with their existing systems. However, software providers will most likely resist incorporating a set of open source address standardization tools into the software they sell to hospital, clinics, and registries because they may have already spent considerable time and effort developing their own closed-source solution for the task. Similarly, open source licensing constraints could potentially dissuade the adoption of such a tool if software vendors using it were required to distribute the remainder of their systems under the same open source license. To overcome these hurdles, it will take the right kind of licensing structures to ensure that commercial software providers continue to develop and release the software needed by cancer registries and other organizations. More importantly, the adoption of a single set of address standardization tools will require firm organizational commitments to maintaining a single consistent level of quality across all consolidated data.

5. References

1. Goldberg, D., A Geocoding Best Practices Guide. North American Association of Cancer Registries. Available at: http://www.naaccr.org/filesystem/pdf/Geocoding_Best_Practices.pdf, 2008.

2. United States Postal Service. Publication 28 – Postal Addressing Standards. Washington, DC United States Postal Service. Available online at: http://pe.usps.com/text/pub28/welcome.htm, 2008.

3. Christen, P. and Willmore, A. and Churches, T. A probabilistic geocoding system utilizing a parcel based address file. In G.J.Williams and S.J. Simoff (Eds.) Data Mining. Berlin, Springer-Verlag: 130-145, 2006.

4. Churches, T. and Christen, P. and Lim, K. and Zhu, J.X. Preparation of name and address data for record linkage using hidden Markov models. BMC Medical Informatics and Decision Making, 2(9). Available at http://www.biomedcentral.com/1472-6947/2/9/, 2002.

5. United States Postal Service. Address Information System Products Technical Guide. Washington, DC United States Postal Service. Available online at: http://ribbs.usps.gov/files/Addressing/PUBS/AIS.pdf, 2008.

6. Winkler, W. E. Matching and Record Linkage. In B. G. Cox et al. (Ed.) Business Survey Methods. New York, John Wiley: 355-384, 1995.

7. Jaro, M.A. Probabilistic linkage of large public health data files. Statistics in Medicine 14(5-7): 491-498, 1995.

8. O'Reagan, R.T. and Saalfeld, A. 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, 1987.

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

10. Michelson, M. and Knoblock, C.A. Semantic Annotation of Unstructured and Ungrammatical Text. In Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05), Edinburgh, Scotland, 2005.

11. United States Census Bureau. Topologically Integrated Geographic Encoding and Referencing System, Washington, DC United States Census Bureau. Available online at: http://www.census.gov/geo/www/tiger, 2009.

12. United States Federal Geographic Data Committee. Street Address Data Standard. Reston, VA United States Federal Geographic Data Committee. Available at: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/street-address/index_html, 2008.

13. Goldberg, D. The USC WebGIS Open Source Geocoding Platform. Los Angeles, CA, University of Southern California GIS Research Laboratory Technical Report No. 11, 2009.

The University of Southern California GIS Research Laboratory seeks to develop cutting edge geographic analysis tools and to apply those tools in ways that increase our knowledge of the built and natural environments while training the next generation of geographic information scientists and promoting the utilization of geographic information science concepts and technologies throughout the academy.

To learn more about our research and teaching programs, contact Leilani Banks, GIS Research Laboratory, University of Southern California, 3620 South Vermont Avenue, Los Angeles, CA 90089-0255

GIS research laboratory http://gislab.usc.edu