The USC WebGIS Open Source Geocoding Platform

Daniel W. Goldberg

GIS research laboratory

Technical Report No. 11

Table of Contents

Executive Summary

This document outlines the current status of the USC WebGIS Open Source Geocoding Platform in terms of the overall goals of the system, its architecture, its data sources, and the technical details of its implementation. Design choices and implementation rationale are provided for each of the components of the system including the geocoding engine, the parsing engine, and the web-based batch processing engine.

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 one of the first steps 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 spatial data sources. 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. The purpose of this document is to provide background information on the current status of the system including the goals it was intended to achieve, its design rationale, and high-level details about its current implementation. As part of this discussion, the details of the parsing engine and the web-based batch processing engine which forms the foundation of the USC GIS Research Lab's suite of online GIS/geospatial processing tools will be described.

2. The Geocoding Process

The USC WebGIS Open Source Geocoding Platform [3] is a postal address geocoding system developed by the USC GIS Research Laboratory. This system implements all of the main components of a traditional geocoding system including address parsing, reference data set definition and storage, feature matching, and feature interpolation. 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 street address is first parsed and normalized to identify standard values for each of the postal address components. After normalization, the system attempts to find one or more reference features that match the input address from within each of the reference data layers that it maintains. 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. A high-level overview of the relationships between each of these components is displayed in Figure 1.

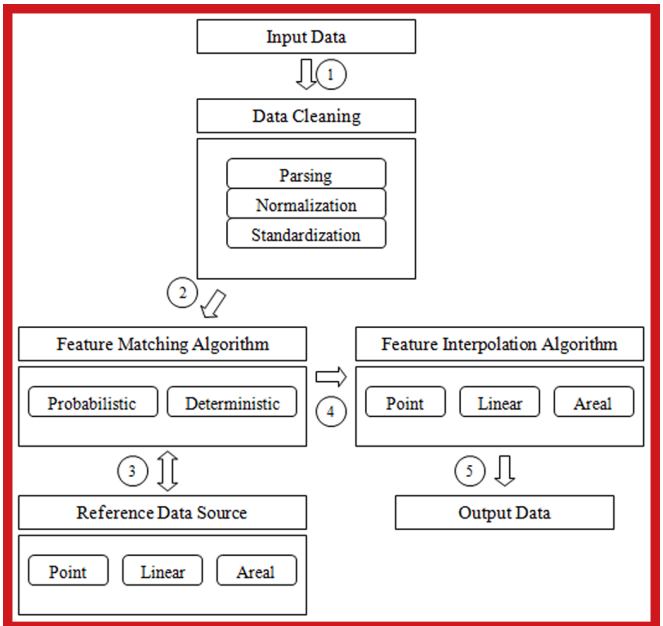

Figure 1 - Geocoding process overview

3. Geocoder Technical Details

As mentioned in the introduction, a geocoder consists of three main processing components (address processing, feature matching, and feature interpolation) and one or more sets of underlying geographic reference data. The following sections will detail the implementation strategy and rationale for each of these components. For each, the rationale for choosing the particular implementation strategy will be

outlined as will be the current status of that implementation. We start, however, by documenting the types of reference data sources used to perform the main processing tasks.

3.1 Reference Data Sources

The data sources used include the freely available 2008 versions of the US Census Bureau TIGER/Lines [8], Places, ZCTA5, ZCTA3, County Subregion, and County files [7], the 2006 version of the Los Angeles County Assessor's Parcel files [6], and the 2009 version of the USPS TIGER/Zip5+4 files [10]. The parcel and Census files were obtained in ESRI shapefile format, while the TIGER/Zip5+4 files were converted to shapefiles from the original flat files. ESRI ArcGIS 9.3 [2] was used to compute the lengths, areas, and bounding boxes of the geographic features after projecting each file from a geographic coordinate system to UTM NAD 1983 Zone 11. The individual TIGER/Zip5+4 segments were aggregated into multi-line features using the ArcGIS dissolve function to create lower levels of resolution TIGER/Zip5+3, TIGER/Zip5+2, TIGER/Zip5+1, TIGER/Zip5, and TIGER/Zip3. Several utilities have been developed to automatically load these reference data sources into their ultimate storage location, Microsoft SQL Server 2008 Spatial. These reference data sources were chosen because (1) they are readily available; (2) they are free or low-cost; and (3) they represent the typical types and qualities of reference data sources routinely used by researchers and organizations for geocoding.

3.2 Address Processing

The input address entered by a user 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 values. 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 [9]. The main goal of an 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").

The implementation strategy was to create 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, all of which are kept in hash tables for quick access. A synonym matching system was implemented that uses these tables 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, it's 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 have already been identified. This implementation is wrapped into a standalone component that can be used in isolation or integrated into other software systems.

3.3 Feature Matching

The feature matching component implemented in the current version of the USC WebGIS Open Source Geocoding Platform is a deterministic matcher that supports attribute relaxation [5], substring matching, and both Soundex and DM-Soundex [11] matching on all attributes. The matching engine accepts a parsed input address which is used to select reference geographic features from spatial databases that contain such things as street segments, parcels boundaries, and address points. If a single match is found for the input address to a reference geographic feature it is added to the set of possible outputs. If no matches are found, the advanced matching techniques are attempted. First, attribute relaxation is employed to try to find less-precise matches to account for errors in the input data or reference data. In this approach attributes of the input address are removed from

the query, first individually then in combination. If a match is not found after fully relaxing all possible combinations, Soundex and DM-Soundex values are substituted for the input values submitted by the user, and the matches are re-attempted using the full exact match as well as attribute relaxation. If the Soundex approaches fail to find a match, substring queries are then attempted by using the SQL "LIKE" clause for each attribute, again first attempting an exact match and subsequently attempting relaxed versions if required. If more than one match is found, a hierarchy-based selection approach is implemented according to the NAACCR standard for all applicable reference datasets [4].

Feature matching can be attempted in either a single- or multi-threaded manner. In single-threaded mode, queries to each of the reference data sources are processed sequentially one after another with each reference data source, first parcel centroids, then parcel geometries, then TIGER/Lines, etc. In the multi-threaded mode, the matches with each reference data source are attempted in parallel, although the queries per-data source are still processed sequentially.

3.4 Feature Interpolation

Depending on the types of the reference features matched, different interpolation algorithms are employed to produce a single geographic point output from the reference feature based on the attributes of the reference features and any other information known about them. The feature interpolation component currently supports address range, uniform lot, and actual lot linear interpolation [1] for street segments using a static 10 m dropback orthogonal to the direction of the street segment. Areal interpolation is performed within SQL Server 2008 Spatial using the built-in OGC STCentroid implementation for all areal unit reference sources, e.g., parcels, cities.
In addition to the desktop interface, the USC GIS Research Lab has also developed a set of web API's that can be used by a user or user-written programs to send address data to the USC GIS Research Lab to be geocoded and returned. This website allows users to access the geocoding functionality along three axes: (1) HTML form interaction, i.e., normal browsing; (2) Web APIs for access from code; and (3) automated batch processing of uploaded data files of varying formats. Each of the underlying geospatial processes are implemented as separate components. The USC WebGIS Online Batch Geo-Processing Engine simply calls the appropriate function within the appropriate library. The web forms and API's make single calls to the components while the batch processing engine spawns an individual thread on the server to start and run the process on the server behind the scenes.

4. Geocoder Output

The output geocode contains the geographic coordinates (latitude and longitude) of the calculated location as well as informational items that provide information on the probable quality of the geocode. The quality of the geocode is expressed as the type of reference feature matched and interpolation strategy used (Table 1), and the type of match obtained (exact, relaxed, soundex, or a combination), and the type of address matched (Table 2).

Table 1 Output geocode quality types

- Exact parcel centroid
- Nearest parcel centroid
- Uniform lot interpolation
- Address range interpolation
- ZIP code centroid
- City centroid
- County subdivision centroid
- County centroid
- State centroid
- Country centroid

In addition, other metadata are returned that shed light on the selection process and criteria used of the reference feature. These metadata include: (1) parsed address which is the result of the address parsing process; (2) the matched address which is the address components that were used in a successful query of the reference data sources; and (3) the reference feature address which is the address information listed

Table 2 Address match types

- Street address
- Post Office box
- Rural route
- Star route
- Highway contract route
- Intersection
- Named place
- Relative direction
- Unmatchable

on the reference feature selected for interpolation. With this information, a user can evaluate the suitability of the geocode using his or her own criteria and/or heuristics.

5. Desktop Geocoding Interface

The USC WebGIS Open Source Geocoding Platform is implemented as a series of independent components that can be re-used independently from each other as well as easily extended as new reference data sources become available and/or new geocoding techniques are invented. A graphical user interface has been developed that allows a user to geocode single records as well as databases of records in batch mode (Figure 2). The desktop client is able to process records locally if the reference data bases are available or it can be used to call a geocoding web service hosted at USC as will be discussed in the next section. The interface allows the user to select which output metadata they would like reported along with their results as well as choose the reference data sources they wish to use (Figure 3) and other geocoding options such as which attributes to allow relaxation on (Figure 4) and which output feature hierarchy to use.

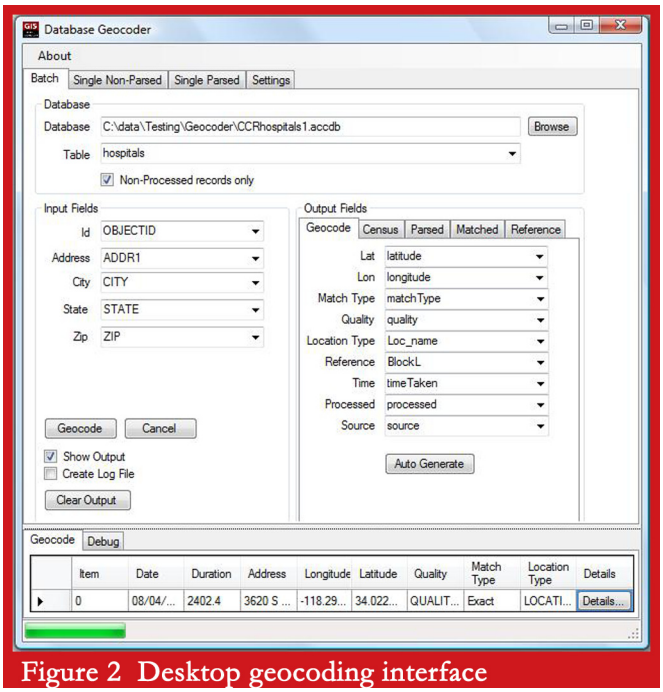

Figure 2 Desktop geocoding interface

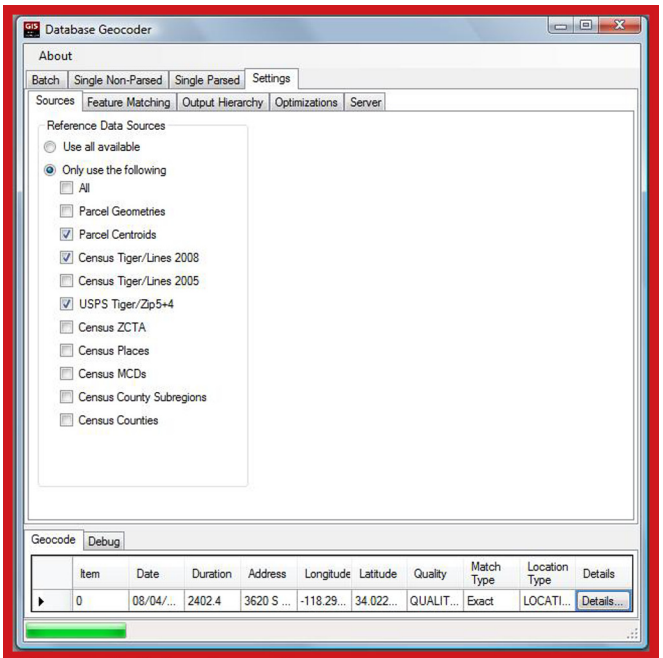

Figure 3 Selecting reference sources

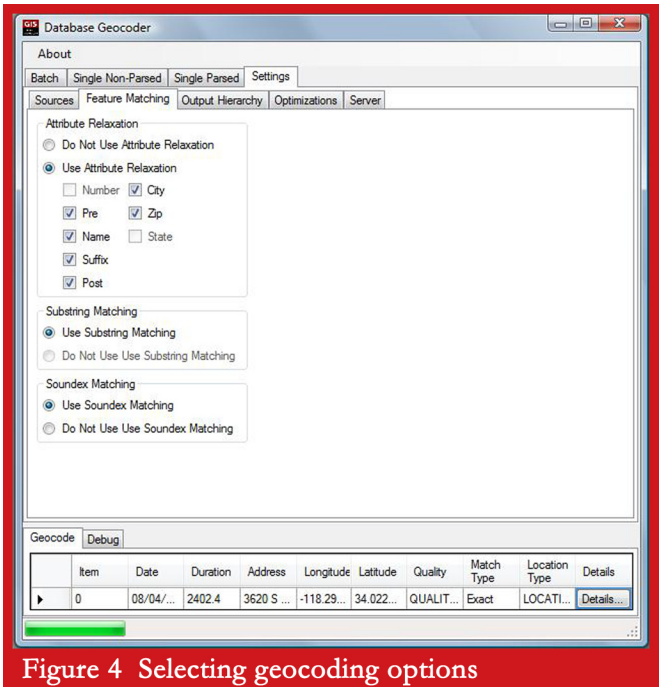

Figure 4 Selecting geocoding options

6. USC WebGIS Online Batch Geo-Processing Engine

6.1 Per-Transaction Model

To ensure that as many people as possible can utilize the service, a threshold system is implemented into the service which limits the amount of processing that any one user can perform at any one time. Users can process as many records (transactions) via any of the three methods (forms, API's, or batch) provided they have a positive transaction balance associated with their account. When users sign up, they automatically get 2,500 transactions. After 2,500 transactions, users can continue to use the service by either signing up as a partner or purchasing transaction credits. To sign up as a partner means they can continue to use the service for free by providing a link to the site on their own web page and providing attribution in any product using or derived from the services. Partners can add 2,500 records at a time to their account once their balance drops below 2,500, i.e., they have a 4,999 transaction limit at any one time. For individuals needing to process large amounts of data at one time, a series of forms allows purchases to be made according to the prescribed payment schedule. These purchased credits can be used in any amount, meaning that they will not be subject to the threshold. All geocode transactions processed by a user from any interface are stored as the user's history in their account on the site.

6.2 Database Uploading

To start a batch process, the user first uploads a data file containing their set of records in Microsoft Access (.accdb, .mdb) or a text delimited file (csv, txt) (Figure 5). This upload occurs from the user's Internet browser (e.g., Internet Explorer, Firefox) over the HTTP protocol using secure socket layer (SSL) encryption. If the file is not an Access file, they must identify the format and layout of the records, after which time a temporary Access file is created and stored on the server to use as temporary storage while the process is running. After uploading, the file is checked for validity, i.e., that it can be opened and its contents read (Figure 6). If valid, the user then maps the fields of the database/table to the correct field of the schema for the process they wish to start (Figure 7). If output fields are not included in the uploaded input database, they are automatically generated and

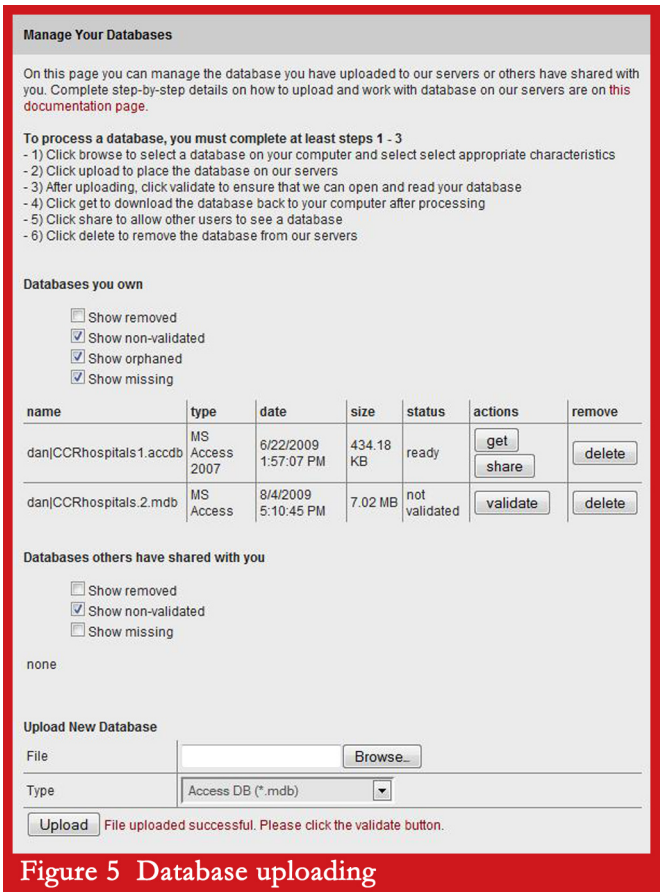

Figure 5 Database uploading

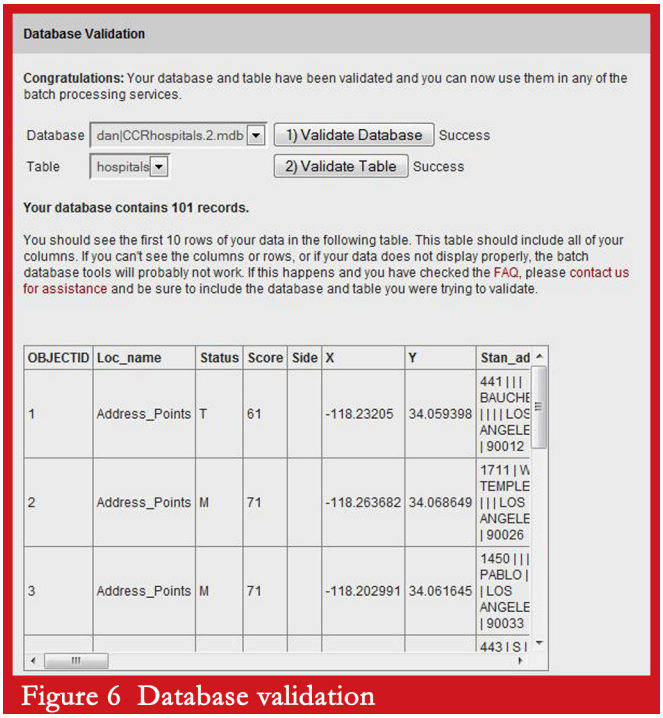

Figure 6 Database validation

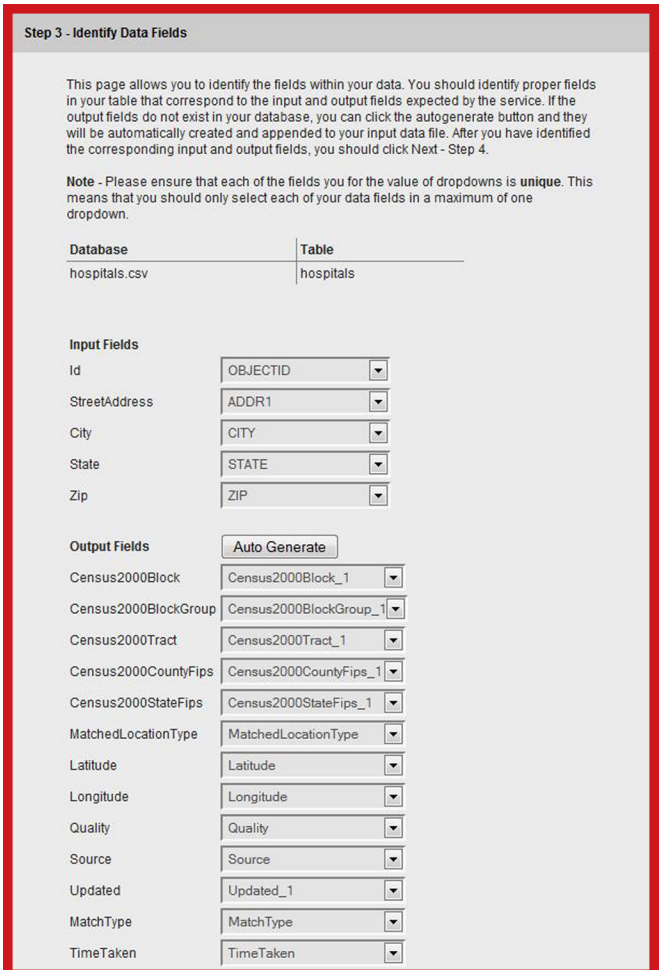

Figure 7 - Database field mapping

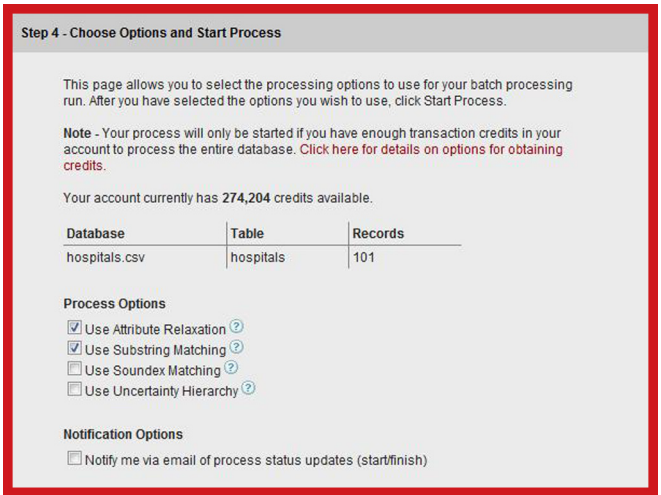

Figure 8 Geocoding options for batch database processing

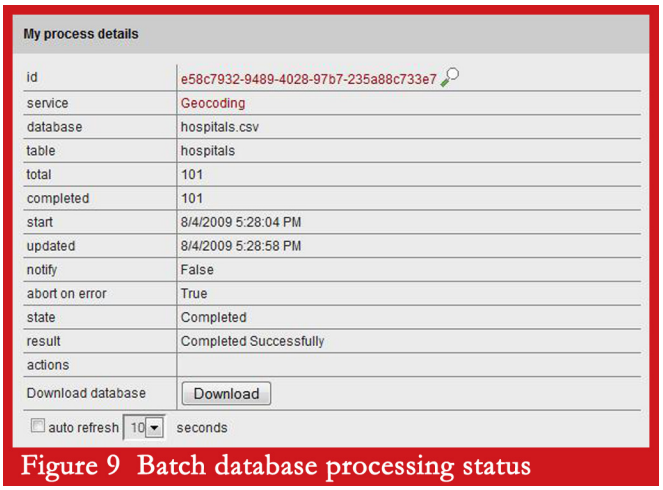

Figure 9 Batch database processing status

appended to it. Uploaded files are automatically deleted from the server after 14 days.

6.3 Batch-Processing Engine

An ApplicationState object is created in the web application memory to manage a set of WebBatchDatabaseRunner objects for each of the services responsible for creating, starting, and stopping the user batch processes. Upon selecting the options for their process and starting the process (Figure 8), the batch processing engine spawns a new thread to run the process and adds it to the WebBatchDatabaseRunner manager in the ApplicationState. The batch processing engine manages each of the threads to ensure it is still alive and processing records. If not, it takes action to resume the thread. When completed (with or without error) the user is notified via email that their process has completed with a link to return to the site to download their processed file (Figure 9).

After downloading the file, the user can then delete the file which removes the data from the USC WebGIS servers.

The USC WebGIS Open Source Geocoder currently supports over 1,600 registered users. To date, the system has geocoded over 7,000,000 addresses in all 50 states. 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.

7. Performance

The USC WebGIS Open Source Geocoder represents the culmination of a multi-year investment into the development of a production-scale, extensible, accurate, and reliable geocoding platform. The current system meets the geocoding needs of many wide-ranging research communities in that it provides a low- to no-cost option for geocoding batches of postal address data using the latest and most accurate freely available reference data sources. The inclusion of several state-of-the-art matching techniques ensure that high-quality results can be obtained and the metadata reported afford consumers a fuller understanding of the true accuracy of each geocode.

8. Summary

Future development efforts will be directed in several key areas. To improve the accuracy of the system, we plan to implement probabilistic matching which will improve both the precision and recall of the system by enabling the production of geocodes that have a high probability of being correct in situations where deterministic matching would otherwise fail to produce any result at all. To improve the usability of the system, we plan to implement methods for making the reference data sources used by the geocoder configurable, instead of the hard-coded way that they are currently defined. This will enable users to utilize their own reference data sources which may be more complete or accurate than what is presently incorporated into the system.

9. References

1. Bakshi, R., Knoblock, C.A. and Thakkar, S. Exploiting online sources to accurately geocode addresses Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems, 2004.

2. Environmental Systems Research Institute. ArcGIS: A Complete Integrated System, 2009.

3. Goldberg, D.W. and Wilson, J.P. The USC WebGIS Geocoding Platform, University of Southern California, Los Angeles, 2009.

4. Hofferkamp, J. and Havener, L. (eds.). Standards for Cancer Registries: Data Standards and Data Dictionary. North American Association of Central Cancer Registries, Springfield, IL, 2008.

5. Levine, N. and Kim, K.E. The spatial location of motor vehicle accidents: A methodology for geocoding intersections. Computers, Environment and Urban Systems, 22 (6). 557-576.

6. Los Angeles County Assessor's Office. GIS ready map base data, 2009.

7. U.S. Census Bureau. Cartographic boundary files, 2009.

8. U.S. Census Bureau. U.S. Census Bureau TIGER/Line, 2009.

9. U.S. Postal Service Publication 28 - Postal Addressing Standards. United States Postal Service, Washington, DC, 2009.

10. U.S. Postal Service. Topological Integrated Geographic Encoding and Reference/ZIP + 4 File, 2009.

11. Wong, W., Liu, W. and Bennamoun, M. Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analystics, Australian Computer Society, Inc., Sydney, Australia, 2006, 83-89.

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

http://gislab.usc.edu