GENERATING VERIFIED AND VALIDATED GEOSPATIAL DATA FROM OPEN-SOURCE WEB 2.0 CONTENT

Barry Bitters, Ph.D., GISP
Leidos, Inc.
Navarre, FL 32566
bittersb@gmail.com

ABSTRACT

This paper describes a semi-automated process developed to extract and merge raw, geospatial feature data from multiple sites on the Internet. This process was tested on varied set of feature classes (specifically airfields, golf courses, maritime lights, police stations, and post offices of South Africa) to evaluate the effectiveness of the procedures. As an illustration, the results for airfield processing are presented. During processing of raw disparate data captured on the Web, attributes were aligned, duplicate records were eliminated, record-level metadata was preserved, and unit of measure conversions were performed. The processed raw data is then merged into a single dataset and attributes are integrated and duplicates are again eliminated. The final quality control review includes a final record-by-record imagery review of all candidate features to insure they actually exist on the ground, their geo-coordinates are accurate and that all duplicate instances have been eliminated. It has been found that this process is faster than the traditional wide-area search technique used to populate and maintain feature databases. The process also provides significant increase in location accuracy and increased feature counts relative to heritage data sources used in the study.

KEYWORDS: Open-Source Data, Public Data, Crowd-Sourced Data, Web-Based Content, Data Mining, Feature Data Capture, Feature Data Extraction, Spatial Database.

INTRODUCTION

A key characteristic of the Web 2.0 environment is the existence of a wealth of open-source and voluntarily generated (crowd-sourced) geospatial data. Produced by a variety of methods and techniques, these geospatial data are presented in varying degrees of reliability, accuracy and precision. Even "authoritative" sources will post geospatial data which for various reasons are often incomplete and erroneous. Numerous forms of “mash-up” and Extraction/Transformation/Loading (ETL) software are currently employed to reuse open-source data. However, these techniques often propagate legacy omissions, duplications and errors from their parent datasets. Publicly available feature databases and gazetteers are often rehashes (with some augmentation) of heritage gazetteer data. Many times data values and name are “stale” and geo-positions are significantly offset. Performing a simple merge of these datasets with newer open-source data cannot provide an accurate, comprehensive inventory of cultural features on the landscape.

Due to incomplete existing databases, the expense of maintaining the currency of existing databases, and the current trend of creating open-source databases using “mashup” techniques, it is difficult to maintain data currency with the temporal nature of human interaction with the built environment and reflect this interaction in accurate and current cartographic, GIS and intelligence inventories/databases. Even those databases produced and made available by government agencies contain omissions, commissions and erroneous data.

This paper presents preliminary results of research into a new approach to repurposing open-source geospatial Internet data for use in augmenting and updating geospatial databases. This approach performs a modified form of logical union of open-source data to arrive at a more comprehensive inventory of cultural landmark features. The primary dilemma addressed is: how best to locate, capture, merge, validate and verify open-source and crowd-sourced geospatial data with an emphasis on insuring the integrity, precision and accuracy of the resulting information.

BACKGROUND

Spatial feature database production is a costly process. Ongoing maintenance of existing feature databases is also a costly endeavor. Even the most comprehensive feature database can never be 100% complete or accurate. Overtime completeness and accuracy will decline in any data store and it is inherent in human-machine interaction that some level of erroneous data will be present. Many organizations will implement a database and due to the high initial cost not ensure the long-term integrity and currency of their data due to the added high cost of a database maintenance program.

Feature databases contain location information and descriptive attributes about natural and cultural features on the ground. Traditionally produced through wide area search techniques, viewing aerial imagery to locate features on the ground, these databases are time consuming and expensive to produce. Maintaining them usually requires the same imagery techniques and the same level of effort as initial production. However, there exists a wealth of open-source geospatial data available on the World Wide Web – all available for viewing and potentially for download. Exploiting these sources of geospatial data could be a low-cost alternative to traditional database maintenance and thereby assist in maintaining currency, accuracy and depth of detail of existing spatial data stores.

Open-source geospatial data comes in many forms. It can be in the form of raw spatial data, text files, spreadsheets, even graphical. More importantly, it can be volunteered data, open-government data, special interest group data, publicly available commercial data or data from research organizations. Each type of open-source data must be viewed from the standpoint of its trustworthiness, accuracy and timeliness. Assuming any open-source data available on the Internet is reliable is a bad assumption, even for data from supposedly "authoritative" providers.

Data conflation (also termed map matching or map merging) involves combining map data from separate sources to create data that is better than either source on its own. A formal definition is "the process of combining geographic information from overlapping sources so as to retain accurate data, minimize redundancy, and reconcile data conflicts" [Lenzerini, 2002]. Attribute validation is a data conflation technique for the confirmation of the correctness of descriptive attributes based on information received from an authoritative source. It is often difficult to find a flawless authoritative source especially when mining geospatial data from the Internet. In the situation where data sources are found not to be authoritative, validation can be achieved, to a limited extent, by applying majority rule techniques to attributes received from multiple sources. This technique involves comparing attributes from different sources and preserving those values that occur most frequently. However, care must be taken when using majority rule techniques since human error can corrupt the most authoritative of sources. Since many Internet data sources have been derived from other existing legacy data, stale and erroneous data is often propagated. Further, when only two legacy sources of data are available majority rule does not apply. During the initial planning phase of this research automated majority rule techniques were explored to resolve differences in attribute data from disparate sources. However, after extensive testing, automated majority rule was found to be ineffective in the attribute conflation process and subsequently abandoned as a means of validating attributes from multiple sources.

METHODOLOGY

A five phase process (Figure 1); Search/Capture, Legacy Data Review, Legacy Data Integration, Final Verification and Output has been designed to transform geospatial data captured from disparate sources into verified and validated geospatial information. In the Search/Capture phase, Web search and Web crawler techniques are used to identify open-source geospatial datasets in the open Web – that part of the Web which is indexed by standard search engines. Specialty searches of the invisible Web are also performed to identify hidden geospatial data. In the

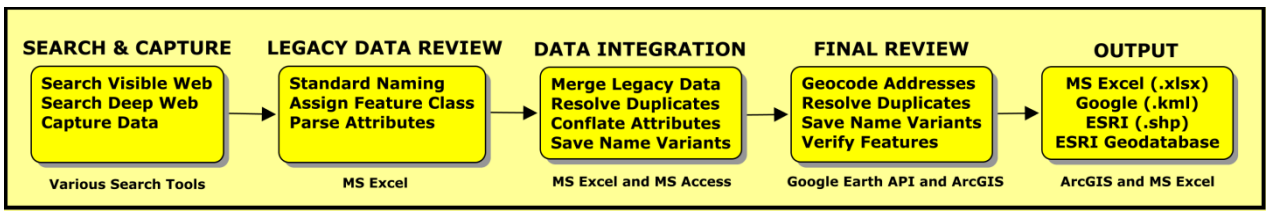

Figure 1. The generalized processing steps used to capture, merge, conflate and verify open-source spatial data obtained from disparate Internet sites.

Legacy Data Review phase, each legacy dataset is reviewed, data fields are parsed, unique alpha-numeric identifier are assigned and record-level metadata is included.

The Data Integration phase is where all captured legacy records are merged into a single database. A specialty merging process is used to perform integration of all legacy data records and attributes. Duplicate record removal of features with the same names is also performed. Resolving duplicate feature records, often termed deduplication, is the process of identifying and resolving duplicate instances of the same feature; and is an essential step when merging disparate data sets. Five types of feature duplication have been identified in both legacy databases and also in databases of merged legacy data.

- Explicitly duplicate names – Those features with the same names at the same location.
- Explicitly duplicate names – Those features with the same name at different locations.
- Nearly duplicate names – Those features with slightly different names at the same location.
- Nearly duplicate names – Those features with slightly different names at different locations.
- Explicitly different names – Those features with different names, but at the same location.

To resolving these situations, potentially duplicate records must be found and then remove in a three step process. The first step in deduplication is concerned with identifying records containing explicitly duplicate and nearly duplicate names and is not concerned with location differences. This process is only concerned with typological aspects. Final duplicate resolution is not performed at this time – only identification and flagging of each record with the identification of its potential duplicate is performed. Duplicate features that are more difficult to identify, those with different names but occupying the same physical location, will be dealt with in later processing.

Final Review phase involves definitive verification of the existence of each final record. Aerial and ground-level image analysis is used as reliable evidence (although temporal in nature) of a features existence. However, rather than depending solely on standard image analysis as confirmation of feature existence, an evidence-based approach has been developed. This approach incorporates a weighted evaluation of both image analysis results and a reliability rating of the open-source data providers to derive a probability of the existence and validity for each final record.

The data verification tasks incorporated in this research were designed to insure all duplicate feature references were eliminated and each feature truly exists in the real world as of the date of the most current imagery available in GoogleEarth™. An integral part of the final verification process is to view each feature on areal imagery and capture a refined geographic coordinate (a six decimal digit latitude and longitude coordinate – datum WGS-84). The verification evidence of feature existence is documented within each feature record.

Feature verification involves formal processes that confirm the existence of a feature and is best achieved through direct visual inspection from the ground – ground-truth. In lieu of ground-truth verification, with the aid of the GoogleEarth™ StreetView™ capability it is now possible in many parts of the world to exploit continuous ground-view imagery along roads and highways. This capability allows the direct viewing of objects along and near to a regions road network. Although this capability allows verification of feature existence, it is not as accurate as ground truth due to the temporal offset of image collection versus image viewing. Care must be taken since the temporal nature of the image capture has a bearing on feature existence and possible change in functionality since the date of imaging.

Image verification is not only confirmation of each features existence, but also a certification of the function of a particular structure. This “eyes-on” image inspection is only possible if current suitable imagery is available. GoogleEarth™ provides the capability to view high-resolution, vertical imagery for much of the world. In addition, GoogleEarth StreetView™ provides the capability to inspect detailed and continuous “hand-held”, terrestrial imagery along highways and in some cases walking paths. This ground view image capability is only available in certain parts of the world. However, in those areas where StreetView™ coverage is available, it can provide indisputable, though time sensitive evidence of feature existence.

This approach has been applied to a variety of different feature types – both cultural features easily recognizable on aerial imagery and features difficult to discern on imagery. During this research, over 5,000 objects have been processed through this validation and verification process with significant increases in feature counts and more accurate geo-positions have resulted.

RESULTS

Table 1 shows the sources of publicly available data used in the airfields portion of this study and includes the number of records captured from each source. At the end of Table 1 is the final record count obtained after conflation and deduplication operations. A 53% increase in airfield records was attained over the record count of the largest legacy data source. Half of the record increase is directly attributable to the logical union of heritage data; while the remaining half were imagery derived during the final image verification step. In summary, we have seen significant feature count increases as opposed to record counts in both authoritative and other legacy sources. Our processes were also applied to other feature classes resulted in a 20% increase in golf course records, a 25% increase in reported maritime lights, a 6% increase in sanctioned police stations and a 12% increase in postal facilities.

A basic assumption during this research has been that no database is 100% complete. Therefore, our results are by no means a comprehensive inventory of **all** occurrences of airfields in South Africa. These results do, however reflect a significant increase in the number of features over what were available in each of the legacy data sources used in the study.

Further, the existence of a major portion of the resulting feature inventories has been verified on either ground-based or aerial imagery. Table 2 lists each feature class analyzed during this project and the size of each final dataset developed. Included in the table are the percentage of records that were actually verified on imagery, the percentage that were viewed could not be verified on imagery and were probably the feature of interest, the percentage that were not actually observed on imagery but were possibly the feature of interest and those features that were unidentified and unverified on imagery. 78.5% of the final merged airfield records were actually verified on imagery, 2.9% were determined to be possible airfields (in most cases abandoned) and 18.6 percent could not be found on imagery (due to inaccurate names and/or geo-locations).

Table 1. Open-source airfield datasets used in this study and their record counts.

Name	*Record Count*
Wikipedia [Wikipedia, 2014]	113
World Aeronautical Data [WorldAeroData.com, 2014]	93
The Airport Guide [The-Airport-Guide.com, 2014]	141
South African Civil Aviation Authority [SACAA, 2014]	350
Great Circle Mapper [Swartz, 2014]	335
World Airport Database [AirportData.com, 2014]	81
Our Airports Database [Megginson, 2014]	435
Global Airport Database [Partow, 2014]	260
ZAF Civil Aviation Authority Voluntary Registration [SACAA, 2014]	37
Runway Finder [AirportNavFinder, 2013]	361
World Airport Codes [Fubra Ltd., 2014]	92
U.S. FAA - South Africa ICAO Location Finder [USFAA, 2014]	264
NGA's Geonames Server [NGA, 2014]	592
Microlighters Forum [Microlighters, 2014]	20
NGA's DAFIF [NGA, 2006]	93
PilotFriend Database [PilotFriend, 2014]	347
Image Derived	259
Final Record Count	**1102**

Table 2. Feature Verification Results

Feature Type	*Final Dataset Size*	*Percent Image Verified*	*Percent Probable*	*Percent Possible*	*Percent Unverified*
Airfields	**1102**	**78.5**	**0.0**	**2.9**	**18.6**
Golf Courses	**577**	**97.9**	**0.0**	**0.6**	**1.5**
Maritime Lights	**229**	**90.9**	**0.0**	**3.0**	**5.7**
Police Stations	**1214**	**55.2**	**20.7**	**14.4**	**9.7**
Post Offices	**2587**	**50.0**	**15.4**	**23.0**	**11.6**

CONCLUSIONS

This research investigated the capture and integration of open-source geospatial data from the Internet. Attributes and geo-positions of a select set of landmark feature classes – airfields, golf courses, fixed maritime navigational lights, police stations and post offices were captured. Airfields and golf courses are easily recognizable on vertical imagery; however, fixed maritime navigational lights, police stations and post offices are more difficult to definitively identify on vertical imagery. Ground imagery in the form of Google Earth™ StreetView™ imagery was used to augment vertical imagery to verify the existence of more difficult feature classes. As a consequence, processing times for each of these classes of features will differ based on an analyst's ability to rapidly discern each feature on imagery.

For easily recognizable feature classes (in this study airfields and golf courses) over a country of approximately 500,000 sq. mi. containing 1000 features of interest, this processing requires approximately one analyst week. For less distinctive features (in this study maritime lights, police stations and post offices) processing times for the same country area and feature count were two analyst weeks or more. This includes the capture and merging of legacy source data, the alignment of attribute data, the capture of naming variations for each record, elimination of duplicate records and the refinement of coordinate values in a semi-automated image review. This timing does not include any image mensuration to capture data to fill empty data fields such as dimensions and other descriptive characteristics nor the research to augment missing data field entries.

However, if an organization is concerned only with the accuracy of location and the elimination of duplicate features; and not concerned with capturing a detailed set of refined and accurate attributes; these estimated processing times would be significantly reduced - by more than half. Because substantial time is required to capture, align, parse and normalize attribute data, eliminating the need to perform these processes would result in a significant time saving. However, it must be understood that detailed attribute information is essential in the identification and removal of duplicate features.

In closing, unlike horseshoes, quoits and A-Bombs, in spatial database production and maintenance, close should not be considered good enough. One basic notion at the beginning of this research has proven to be totally false. It is not a reliable expectation to assume a simple logical union of data from disparate open-source datasets will provide reliable information. Merely applying Extraction, Transformation, and Loading (ETL) techniques or "mash-ups" to Internet derived spatial data will not produce accurate, reliable and current feature data. Considering the flawed general impression that data published on the Internet is correct, every effort should be taken to insure the accuracy of legacy spatial data reused for other purposes. Prior to the reuse of any legacy data a formal validation and verification process is essential to insure that "garbage in, garbage out" is not the result.

Our approach to the validation and verification of publicly available geospatial data incorporates a rigorous review and comparison of data from multiple sources – both legacy data sources and imagery sources. This process is designed to capture all recorded feature instances, remove duplicate instances, preserve the most accurate attribute values, save naming variations and refine geo-positions. As the results above indicate, exploiting publicly available Internet data does provide added value in the form of quantity, quality and accuracy. Although the processes used during this research are somewhat time-consuming, they do provide significant increases in feature counts and geo-positional accuracy over their parent legacy feature data available on the Internet.

During this project, minimal effort has been directed toward software development. However, by refining the software used in the effort, we feel that a significant saving in processing time can be achieved. Future efforts will be directed toward integrating and refining software to streamline processing and thus reduce the production timeline. With additional automation of tasks and the resulting reduction in processing time, this technique of generating verified and validated spatial data could provide a cost effective means to generate or augment geospatial databases for use in a wide range of location-based functions.

REFERENCES

AirportNavFinder, 2013. *Runway Finder*. http://airportnavfinder.com/. Last viewed 01 Feb 2013. Site no longer available.

Fubra Ltd., 2014. *World Airport Codes*. http://www.world-airport-codes.com/. Last viewed: 15 Feb 2014.

Lenzerini, M., 2002. Data integration: A theoretical perspective. Proceedings of the twenty-first *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. Madison, WI, USA, 02-06 June 2002; ACM New York, NY, USA.

Megginson Technologies, 2014. *OurAirports Database*. http://www.ourairports.com/. Last viewed: 15 Feb 2014.

Microlighters, 2014. *Microlighters Forum*, http://microlighters.co.za/. Last viewed: 15 Feb 2014.

NGA, 2014. *NGA's Geonames Server*. http://earth-info.nga.mil/. Last viewed: 15 Feb 2014.

NGA, 2006. *Digital Aeronautical Flight Information File (DAFIF).* From personal archive, data no longer available on open Internet.

Partow, Arash, 2014. *Global Airport Database* http://www.partow.net/. Last viewed: 15 Feb 2014.

PilotFriend, 2014. *PilotFriend World Airfield Database*. http://www.pilotfriend.com/. Last viewed: 15 Feb 2014.

SACAA, 2014. *South African Civil Aviation Authority*. http://www.caa.co.za/. Last viewed: 15 Feb 2014.

Swartz, Karl L., 2014. *Great Circle Mapper*. http://www.gcmap.com/. Last viewed: 15 Feb 2014.

The-Airport-Guide.com, 2014. *The Airport Guide*. http://www.the-airport-guide.com/. Last viewed: 15 Feb 2014.

USFAA, 2014. *South Africa ICAO Location Finder.* https://www.notams.faa.gov/. Last viewed: 15 Feb 2014.

Wikipedia, 2014. *List of airports in South Africa.* http://en.wikipedia.org/wiki/List_of_airports_in_South_Africa. Last viewed: 15 Feb 2014.

WorldAeroData.com, 2014. *World Aeronautical Database* http://www.worldaerodata.com/. Last viewed: 15 Feb 2014.

AirportData.com, 2014. *World Airport Database*. http://www.airport-data.com/. Last viewed: 15 Feb 2014.