Developing Open source based tools for geospatial integration from disparate sources

Heshan Du[1], Suchith Anand[1], Jeremy Morley[1], Glen Hart[2], Didier Leibovici[1], Mike Jackson[1]

[1]Centre for Geospatial Science, University of Nottingham, UK

[2]Ordnance Survey, UK

Tel: +44(0)115 846 8411

Email: psydhd@nottingham.ac.uk ; URL: http://www.nottingham.ac.uk/cgs/

1. Introduction

The context of this paper is the progress of national and international spatial data infrastructures, such as UK location Programme and INSPIRE, contrasted against crowd-sourced geospatial databases, such as OpenStreetMap (Anand et al, 2010). Crowd-sourced data sources rely on volunteers to collect data. Though it is not as structured as authoritative data, crowd-source data may provide rich source of information and updates more frequently and also includes many interesting user-based information. Though currently being relatively independent, authoritative and crowd-sourced communities need to communicate and collaborate to improve the overall quality (richness, consistency, accuracy, and timeliness) of geospatial information. It is desirable but challenging to generate an overview of all available information of any object, from disparate sources, with differing conceptual, contextual and topographical representations. Furthermore, in the ever-changing world, there is an increasing need for the representation of knowledge of objects to be fluent, changing during its use (Bundy, 2006).

Ontology, in information science, refers to a formal representation of knowledge by a set of concepts and their relationships within a domain. It is hailed as a mechanism to make better use of the Web, by offering a shared definition of a domain that computers can understand enough to meaningfully process data automatically. Ontology is expected to play a major role in the Semantic Web, which is to add a level of meaning to the Web (Wilson, 2004). Geo-ontology, as a sub concept of it, refers to the formalization of concepts sharing among GIS field (Yang et al, 2006), resulting from analysis and modelling of ontology in geo-spatial application (Wang et al, 2007). Describing the characteristic of data and resource and data acquiring mode, geo-ontology can provide a uniform expression for data integration, sharing, and updating.

This research has been carried out to understand the issues of data integration between crowd-sourced information and authoritative data. ***Geospatial data integration*** (GDI) in this context refers to combining geographic data, including spatial and non-spatial data, from disparate sources, with differing conceptual, contextual and topographical representations. Ordnance Survey (OS) Integrated Transport Network (ITN) data and OpenStreetMap (OSM) road data for Portsmouth, UK were used as test data for case studies.

2. Ontology Plugin for QGIS

The first stage of this project involved looking to existing open source GIS packages and extending their functionality for this research question. We developed new functions as plug-ins of QGIS, which is a popular Open Source GIS application providing data visualization, editing, and analysis capabilities (qgis.org, 2010). The first main research question is how do we define data from different sources are corresponding, referring to the same feature. The following basic relationships for the features in the crowd sourced and authoritative data are considered

SamePlace(featureA, featureB): featureA and featureB are in the *same* place [e.g. *Portsmouth, UK*]

Near (featureA, featureB, *m*): featureB is within *m* metre buffer of featureA

SameName (featureA, featureB): featureA and featureB have a *same* name. Firstly a list with necessary definitions, such as "ST = Saint", is kept. In addition, for comparing the name strings, an edit distance is defined.

SameCategory (featureA, featureB): featureA and featureB are of a shared category.

Neighbour (featureA, featureB, m): featureA and featureB have at least one point in common, given *m* metre fuzzy tolerance.

SameNeighbour (featureA, featureB): featureA and featureB have *at least one* neighbour with a *same* name.

Based on the definition above, SameFeature is defined as following:

SameFeature(featureA and featureB) =

SamePlace/*Near*/\SameName/\SameCategory/\SameNeighbour

One of the key problems for implementing matching algorithm based on the above concept, is that there no information on neighbours stored in one of the datasets used (OSM data). To solve this problem, a Network Building algorithm is implemented.

Also there are key integration issues with incomplete dataset (in OSM data some required information is incomplete; for example some fields show name=NULL, OSM: highway=unclassified or NULL). To handle this problem, a probability based approach is taken. For example, for a named road, *SameFeature=SamePlace/\Near/\SameName/\SameNeighbour*.

The formula used for calculating the probability of two features being the same is defined as following:

Probability (SameFeature) =

w1*samePlace+w2*near+w3*sameName_Category+w4*sameNeighbour;

w1+w2+w3+w4=1

A user-friendly interface is designed to allow users to input directly whether two data sets describe the same place, and to specify the buffer size to define near and the fuzzy tolerance for the concept neighbour, as well as weights (w1， w2, and w3) to put on these constraints.

The feature matching algorithm was designed and software based on the methodology described above, is developed as a plugin in QGIS. Its graphic user interface is shown below (Figure 1). It requires users to specify two input line layers, and some corresponding fields, such as name fields in both data sets. It enables users to conduct different experiments easily, by specifying fuzzy concepts, such as buffer size, and assigning different weights in score function variables.

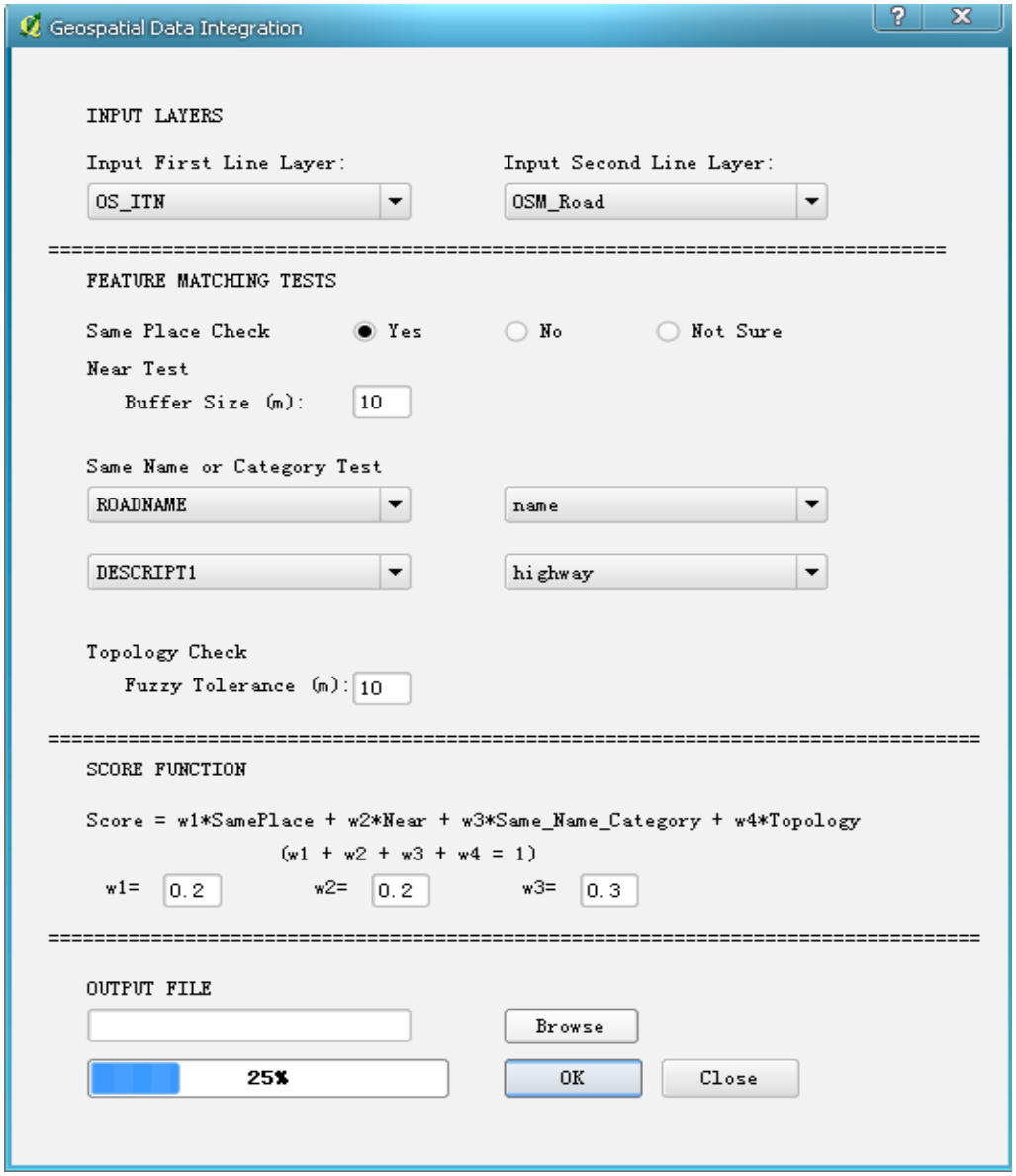

Figure 1: GDI tool developed

3. Standalone open-source software development for geospatial integration

The second stage of the project looked into how to provide high quality geospatial information to users taking the best of available information. For this, the research focuses on two main issues:

1) How to link information from disparate sources
2) How to merge the linked information.

The first problem is similar to the problem of finding correspondence in our previous research. However, this time, we tackle it using a different approach, which is through translating to a uniform ontology written in OWL. In doing so, some tolerance is often required. When merging linked information, it is also necessary to guarantee the merged information is consistent. Consistency means there are no logical conflicts and it should follow some common sense rules. Ideally, the geometries of a same object in disparate datasets should be the same. If there is any inconsistency, some selection or amendments are often required based on the accuracy and timeliness of information. To start with, it

enables users to select between two geometries.

The methodology used is based on ontology, which refers to a collection of information in a particular domain. OWL, a W3C standard web ontology language is used to represent ontology (www.w3.org, 2010). Pellet, a theorem prover and OWL reasoned, is used to discovery inconsistency (clarkparsia.com, 2010). To solve the above problems for road networks, a graph model is employed. Within this model, road network is simplified as a graph, made of edges and vertexes. Each road is seen as an individual of base class edge, while each end point of road is represented as an individual of base class vertex. Information, including attributes and geometry, about each individual is stored as data properties, and relationships between different individuals are represented as object properties.

The prototype software (Figure 2) was implemented employing the methodology and algorithm above using Java. It imports a Java API developed by JUMP. JUMP Unified Mapping Platform is a Java open source application for viewing, editing, and processing geospatial data (www.vividsolutions.com, 2006). It allows users to import OWL or shapefile data (It will translate shapefile data into OWL data) and visualizes OWL data as a graph. When a road is searched, both attributes and geometry information about it will be shown. For instance, Figure 2 shows what happens when Adstone Lane is searched, using Ordnance Survey ITN data and OpenStreetMap data of Portsmouth imported. In addition, it allows users to merge information from different datasets, and store it using OWL. If the newly generated ontology is inconsistent, it enables users to select between two geometries by indicating preference degrees (Figure 3).

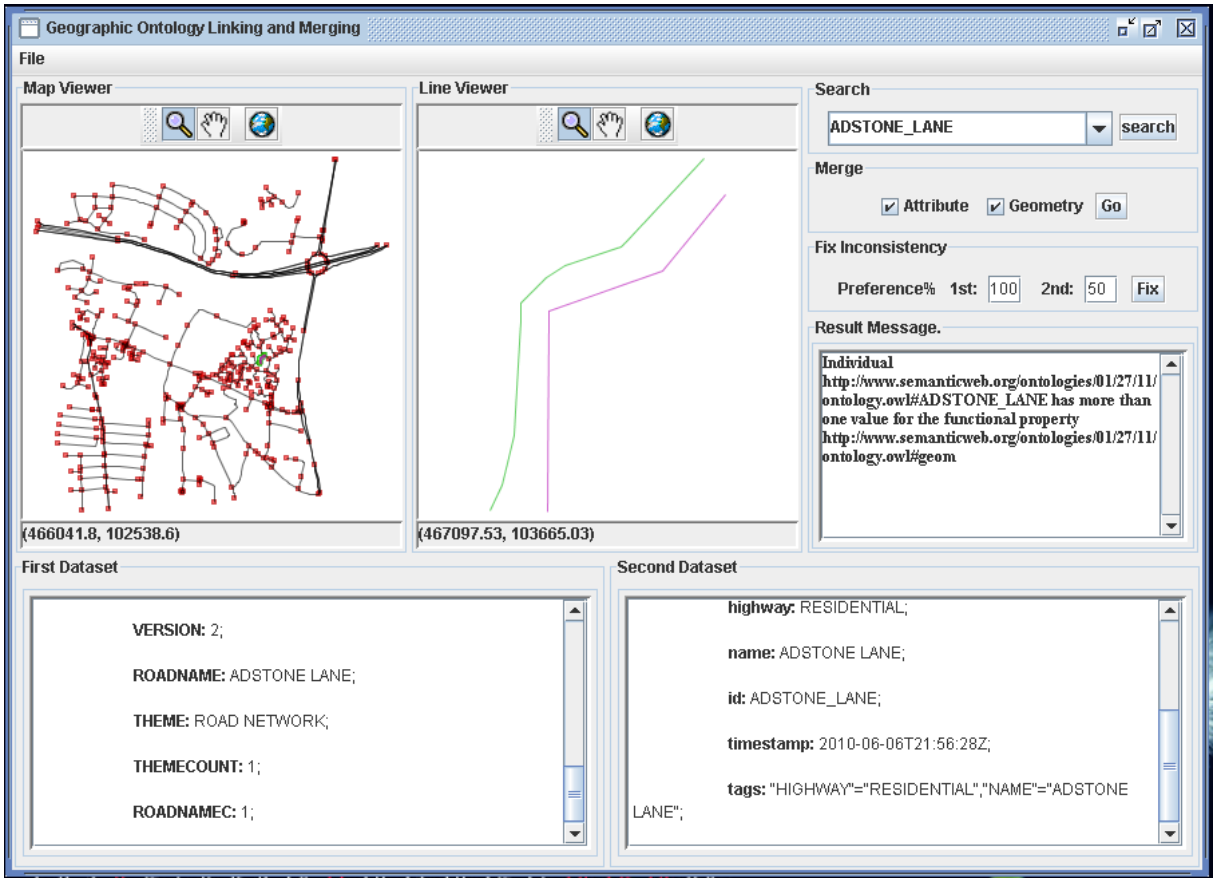

Figure 2: GDI open source software: import

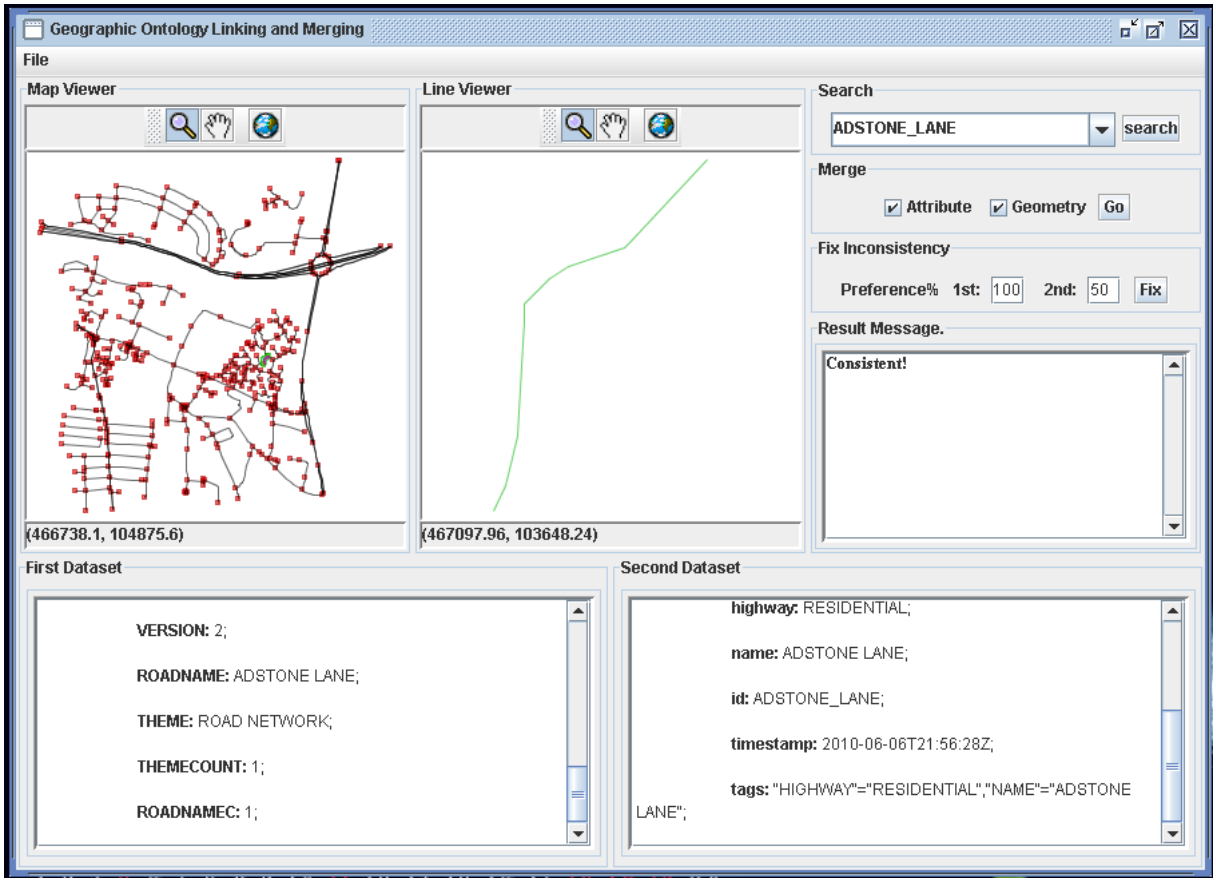

Figure 3: GDI open source software: selection

4. Conclusions and Future work

This paper summarizes our research on geographical information integration and related open source software development. Ontology based methodologies were developed and implemented to aid users integrate geospatial information from disparate sources. Firstly, we developed new functions as plug-ins of QGIS, a popular open source GIS software tool, using Python. Then we moved on further to develop a standalone open source software versions using JAVA to tackle the challenge of geospatial information linking, merging and updates. The results are promising but more work need to be done in refining the process in linking information and for inconsistency resolution. Future work will concentrate on developing more reasonable strategies for inconsistency resolution to solve different real life problems

5. Acknowledgements

The authors express thanks for the Ordnance Survey, UK and OSM for the data used in this work. All figures in this text using OS data are ©Crown Copyright/database right 2010. An Ordnance Survey/EDINA supplied service.

6. References

Anand, S., Morley, J., Jiang, W., Du, H., Hart, G. and Jackson, M. (2010). *When worlds collide: Combining Ordnance Survey and OSM data.* AGI Geocommunity Conference

Association for Geographic Information (2010). *AGI Foresight Study: The UK Geospatial Industry in 2015*. [Online] Available at:

http://www.agi.org.uk/storage/AGI%20Foresight%20Study%20Summary%20Report%201.1.pdf .

Bizer, C., Heath, T. and Berners-Lee, T., (2009). *Linked Data-The Story So Far*. [online] Available at: http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf .

Bundy, A., & McNeill, F. (2006). *Representation as a fluent: An AI challenge for the next half century*. IEEE Intelligent Systems.

Clarkparsia (2010). *Pellet: OWL 2 Reasoner for Java* [online] Available at: http://clarkparsia.com/pellet [6 Oct 2010].

Du H., Jiang W., Anand S., Morley J., Hart G., Jackson M. (2010). 'Ontology Based Approach for Geospatial Data Integration'. In International Cartography Conference 2011, Paris. (submitted).

Quantum GIS 2010, *Welcome to the Quantum GIS Project* [online] Available at: <qgis.org>. [7 August 2010].

Vividsolutions (2006). *JUMP Unified Mapping Platform* [online] Available at: <http://www.vividsolutions.com/jump> [8 Oct 2010].

Wang, Y., et al (2007) *Geo-ontology Design and its Logic Reasoning*, Geoinformatics 2007: Geospatial Information Science.

Wilson, R (2004) *The Role of Ontologies in Teaching and Learning*. TechWatch Reports, Citeseer, 2004.

W3C (2004). *OWL Web Ontology Language Overview* [online] Available at: http://www.w3.org/TR/owl-features [6 Oct 2010].

Yang, K., et al (2006). *The Research and Practice of Geo-Ontology Construction*. Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion.