Proceedings of the International Multiconference on
Computer Science and Information Technology, pp. 305 – 311

ISBN 978-83-60810-22-4
ISSN 1896-7094

Language and Location: Map Annotation Project – A GIS-Based Infrastructure for Linguistics Information Management

Yichun Xie
Institute for Geospatial
Research and Education,
Eastern Michigan University,
Ypsilanti, Michigan 48197, USA
(Tel: 001-734-487-7588; email: yxie@emich.edu)

Helen Aristar-Dry, Anthony Aristar,
Hunter Lockwood, Josh Thompson,
Dan Parker, and Ben Cool
Institute for Language
Information and Technology,
Eastern Michigan University,
Ypsilanti, Michigan 48197, USA
(email: hdry@linguistlist.org;
anthony.aristar@gmail.com)

***Abstract*—The Language and Location: Map Annotation Project (LL-MAP) has been funded by the US National Science Foundation to build a database of linguistic information integrated into a Web-based geographical information system. LL-MAP embodies several innovative concepts of computational linguistics, such as spatial data engine driven architecture, dynamic joining of linguistic information with related cultural and geographic data, multi-layered and linked visualization, real time online data harvesting, collaborative toolboxes for linguistic studies, quick search of digital gazetteers, and toponymical analysis. This paper will demonstrate these LL-MAP functions and discuss their disciplinary implications in linguistic studies.**

***Keywords*—computational linguistics, geographic information system, geography, web-mapping, digital gazetteer**

I. Introduction

Computational Linguistics (CL) has traditionally concentrated on natural language processing and the manipulation of text corpora, e.g., by parsing, by statistical analysis, and by data mining. However, CL is now being extended to encompass new information structures and the queries that address them. One such structure is the digital gazetteer (DG) and its associated systems of information exchange. DGs play an important role in geographically enabled information management and retrieval systems. Not only do they contain critical relationships between place names and geographic locations but they also enable linking between these elements of place description and information on the languages and cultures of the region (Goodchild and Hill, 2008). The latter function requires the participation of ancillary systems, such as the LL-MAP system which is the subject of this presentation. LL-MAP (Language and Location: A Map Annotation Project, http://LL-MAP.org), is a 3-year project, funded by the National Science Foundation of the United States, which is designed to create a geographically-enabled language information system which is modeled on the Digital Gazetteer Information Exchange (DGIE) system.

Linking between geographic coordinates and a place name constitutes a direct linking within a DGIE system. Indirect linking, in turn, occurs when cultural, demographic, or historical information is linked to the geographical coordinates using the place name as a key. Indirect linking fuels a vision in which literary texts, library catalogues, archeological records, or any other source which mentions a place may be addressed with complex queries, such as, "Find all festivals which occur in the area where this language is spoken." In the development of such a system, The Alexandria Digital Library (ADL) project at the University of California, Santa Barbara is a pioneer project, as is the ECAI project at the University of California, Berkeley (see: http://www.ecai.org/projects/gazetteer/nsf_multisys_proposal.html).

The LL-MAP project is not itself a digital gazetteer; but it is a comprehensive source of georeferenced language information which can participate in a DGIE. Moreover, it resembles a DGIE in that it (a) effects a direct linking between a language and the geographical coordinates of the areas in which it is spoken and also (b) uses the language (or rather the ISO 639-3 code for the language) to effect indirect linking between the place and a wealth of resources related to the language.

The field of linguistics is just beginning to explore geographically enabled information management; but it represents a natural pairing. Geography, like language, is an expedient means of organizing large amounts of disparate data: just as almost any event, observation, or object can be related to a place, so can it be associated with a language. That is, place names and their geospatial locations can serve as the digital keys for organizing linguistic information and its relationships with the geographic and social factors of the place (Jones et al, 2008; Guo, Liu, and Wieczorek, 2008). LL-MAP exploits this association to promote new data juxtapositions and hence the generation of new knowledge.

New linguistic insights with cross-disciplinary application can best be realized in a system that melds language information with information from the physical, geographical, and social sciences. The most effective way to do this is through a Geographical Information System (GIS), which can flexibly organize a wide range of het-

erogeneous data, integrating language data with geographical, political, demographic, zoological, botanical and archaeological data in ways which are immediately visually interpretable. This is the motivation for the development of the LL-MAP project. The LL-MAP project is designed to build a database of linguistic information which is integrated into a geographically-referenced system. The LL-MAP system allows users to generate customized maps showing the relationships between language and diverse kinds of non-linguistic data. This system has collected most available online linguistic information resources and developed several toolsets, which allow researchers to harvest digital map services from other publically assessable online resources. The integrated data approach embodied in the LL-MAP system also promotes innovative research methods. For instance, the LL-MAP system supports toponymical analyses that are important in allowing field linguists to understand and visualize both physical and cultural landscapes pertaining to languages and give insights into the language evolution and its relationships to political, economic and cultural histories of a region. The system also enables researchers to incorporate temporal and spatial information connected by relationships inferred from historical interpretation, and the visualization of these data through linked maps and timelines. Moreover, the collaborative research toolbox of LL-MAP supports adding annotations to map-oriented data, and discussing the relationships the system manifests. In this way LL-MAP encourages collaboration between linguists, historians, archaeologists, ethnographers and geneticists, as they explore the relationships between language and cultural adaptation and change.

II. Spatial Data Engine Driven Language Database Management System

The LLMAP project team surveyed numerous linguistic, non-linguistic and geographic data sources available on the Web and in the partners' collections. A rich collection of data has been processed and integrated into the LLMAP system, having been prioritized on the basis of the data's relevance to project goals and the tractability of the data sources. The language data were initially collected from Global Mapping International, the partnering organizations and researchers, US population census data, and online tractable sources. Non-linguistic global data sets consist of 1) Geography (including, Ecoregions, Rivers, Lakes, Man-made Waterways); 2) Political Divisions (including, Country, States / Provinces, Counties, Census Tracts (U.S. only), Major Cities, Smaller Populated Places); 3) Transportation (Roads, Railways, and Airports). Some additional data on flora, fauna, and population demographics are also viewable as tables and made available for real-time mapping to base layers. These features are particularly useful in research involving language development, language contact, and language spread.

To these datasets have been added geodata created by vectorizing paper maps published in a number of important modern and historical language atlases. Over 280 published language maps have been scanned and geo-registered, and 85% of these have been integrated into the LL-MAP user interface for public viewing. For reasons of copyright, the other 15% are currently viewable only on the development site. Ninety per cent of the maps on the interface present language-related information, and 10% of the maps show data from the physical and social sciences.

The spatial data engine (SDE) based geo-database approach was adopted as the key structure of the LLMAP Web Portal (Fig. 1). The LLMAP Geodatabase integrated language data and language-related physical and social information, either spatial or non-spatial. The Geodatabase was built on the Oracle 10g relational database and ESRI's ArcSDE 9.2. This database is very extensible. New data, either attribute tables or spatial layers, can be easily imported to the database by the customized Client tools, and the imported data can be quickly published on the website.

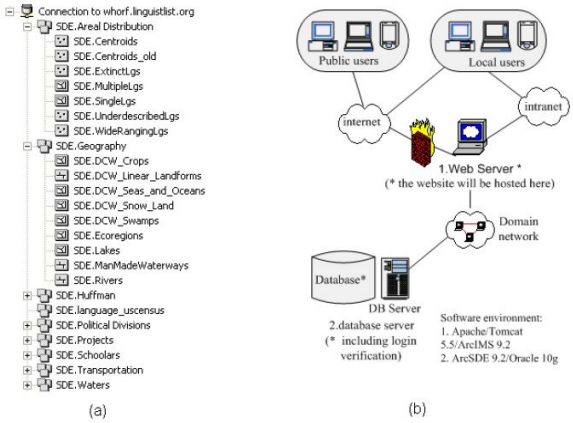

Fig. 1 LLMAP Database and System Architecture - (a): a subset of data in LLMAP Geo-database; (b) the system architecture of LLMAP

III. Quick-Linked Language Information Retrieval

On the basis of the aforementioned SDE geographic referencing and a standard language code, the LLMAP project created a "Quick Map" facility, which provides quick and easy access to language data through search by either (a) language name, (b) country and region or (c) language family. Quick Map is integrated with a facility called the Data

Browser, which can access all the data in the LINGUIST List web of interlinked databases of language information (http://linguistlist.org). Because all languages have a unique ISO 639-3 code and all LINGUIST List data are classified by that code, the Quick Map facility can find all the data on that language in the facilities on our site. The data include the NSF projects MultiTree (a database of language relationship data) and EMELD (a website and database on endangered languages and standards related to their preservation). Thus language trees containing a language, linguists who work on that language, books that deal with it, reviews of those books and dissertations on the language — all these are visible in the Data Browser. Archival data linked to that language in the OLAC (Open Language Archives Community, http://lan-

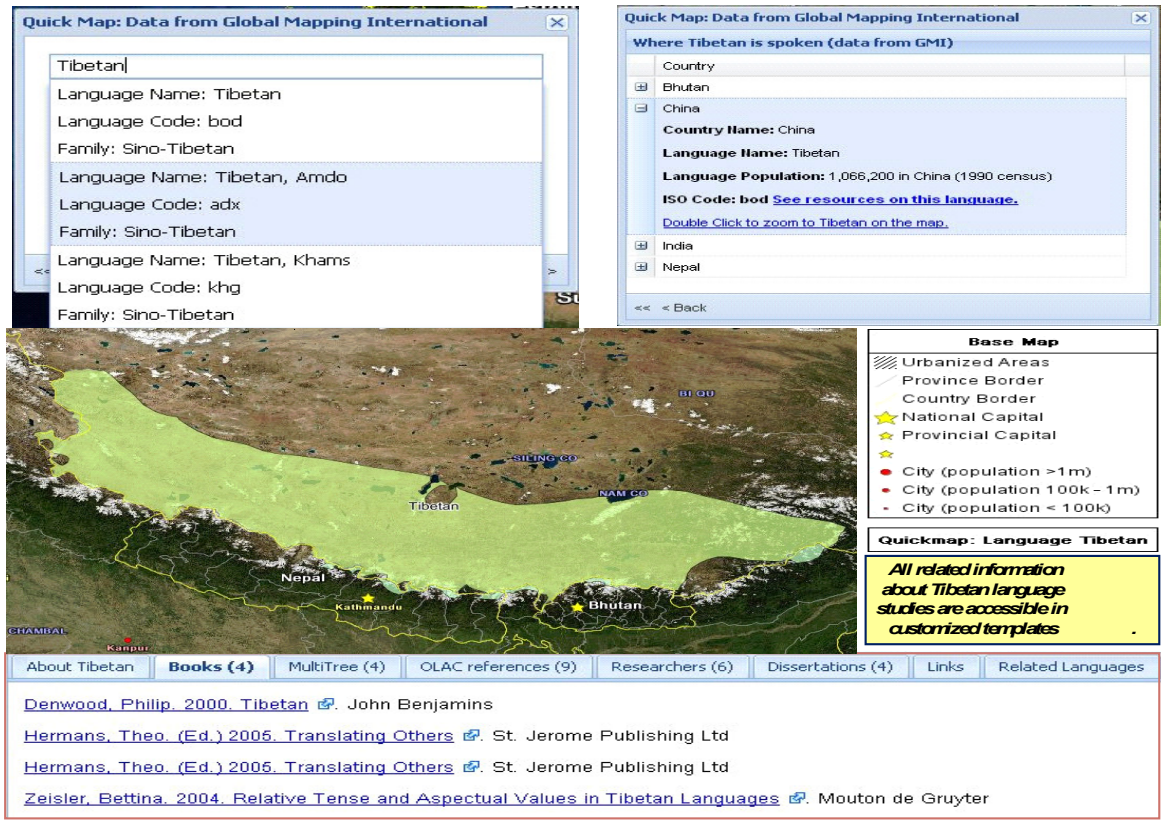

Fig. 2 A Linked View of Quick Language Query

guage-archives.org) archives are also indicated by a link. A screenshot of a search by a language name (Tibetan) is provided in Figure 2.

Text-based search through geo-parsing for deriving spatial locations and navigation services is very important to geographically enabled Web mapping systems (Jones et al, 2008; Guo, Liu, and Wieczorek, 2008). LLMAP adopted the relational referencing method for mapping both spatial and non-spatial features (languages in the context of LLMAP). Mapping languages often involves a complicated process to locate material in a precisely geographical context. A great deal of data either does not contain spatial information or has only a very general description of the location of speakers of a language. For example, the US census data is a valuable source that can be used to study social issues of languages. Unfortunately, the Census data source has no spatial information, except for one field that indicates where particular languages are spoken. A common solution in GIS terminology would be to join the census data with a spatial data layer, e.g. a US state layer, a US county layer and a US tract layer, and thus map the three layers with language information appended. In reality, it is almost impossible to foresee what types of maps a user wants to create and preload them because the LLMAP system aims at providing capability for real time and dynamic searching and mapping language information. Thus, we have to use innovative ways of joining different sets of data and designing maps in real time so that we can spatially locate phenomena on maps. To solve this problem, we designed a dynamic joining method on the basis of combining the attribute-based joining with geographical location-based joining to generate any spatial layer on the fly called an acetate layer (http://www.esri.com/news/arcuser/0102/files/tutorial3.pdf; Fig. 3).

IV. Web-Data Harvesting

Online data harvesting is an important feature of Web-based digital library (Yong, 2004). As we will discuss in the next section, the LLMAP project is creating a transparent and Web-enabled data-sharing and research-collaboration system. Therefore the ability to harvest data from the Web is a critical tool for LLMAP. This data harvesting is built on the basis of the OGC WMS implementation specification (http://www.opengeospatial.org/standards/wms). The LLMAP system has a considerably advanced WMS harvesting tool. Although other systems allow the material from WMS servers to be harvested, none allows them to be harvested so easily. As Figure 4 illustrates, if one goes to the LL-MAP interface (http://llmap.org) and selects “Enter Layer URL,” the system will allow the user to insert the GetCapabilities URL of a WMS server. When this is done, the system will extract the layers that the WMS server is making available, and allow users to select which of these they wish to insert on the map. The transparency of the layer is also selectable. This is important, for if layers are to be seen one on another, they must be transparent enough to allow all layers to show through. The end result is a map that can show many layers at once, each layer displaying different kinds of information.

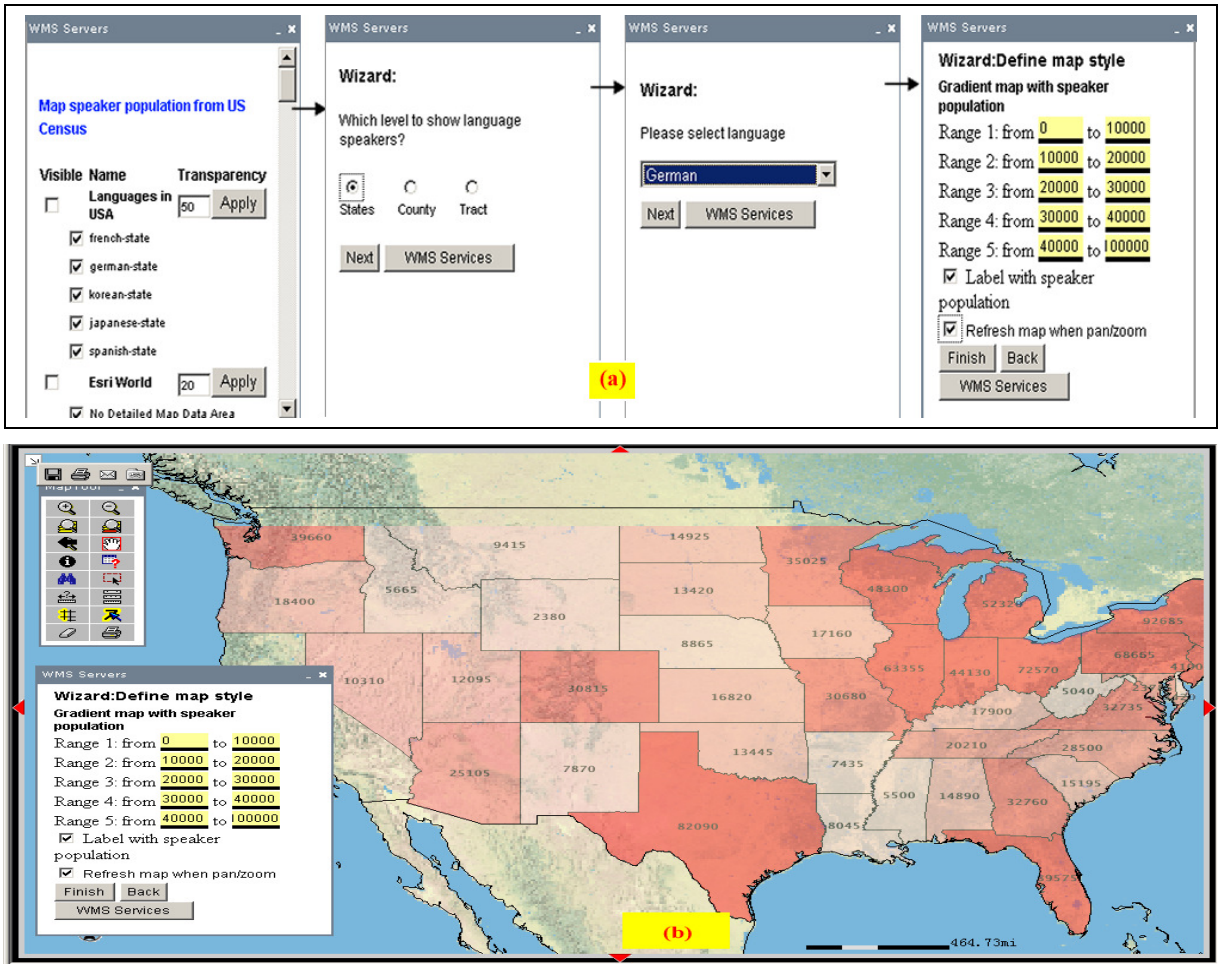

Fig. 3 Dynamic joining wizard (a) and the generated acetate layer labeled by language speaker population with gradient color (b)

The system thus allows many layers of information to be inserted on the fly onto a language map of the user's choice. This is a very useful tool in circumstances where a user wishes to assess the context in which a language or a subset of vocabulary items exists

V. Toponymical Analysis

Through multiple-layered and linked viewers (including maps, graphics, multi-media, and templates of real-time extracted descriptions / annotations), the LLMAP system supports toponymical analysis, which can link and visualize both physical and cultural landscapes pertaining to languages and give insights into the language evolution and its relationships to political, economic and cultural histories of a region. As we know that majority of toponyms are derived from identifiable features of both the natural and manmade landscape, including water sources, landforms, bioforms, and passageways. It is commonly agreed that toponyms serve as an environmental record of indigenous knowledge. Therefore, scientific visualization of the relationships between toponyms and the natural and manmade landscapes will help interpret and understand the meanings of toponyms (Fig. 5).

VI. Temporal and Spatial Analysis of Language Changes

Language change at a place has been closely associated with the history of that region, including dynamics in culture, economy, environment, politics and the entire society. It is critical for a digital linguistic information system to incorporate temporal and spatial information connected by relationships inferred from historical interpretation, and the visualization of these data through linked maps and timelines. The LLMAP portal is designed to accommodate these types of analytical functions for visualization and annotation. For example, Figure 6 presents the migration of Celtic-speaking populations between 6th century BC and 3rd century BC, and the subsequent expansion of the Celtic territory. Additionally, the map traces the influence of the Hallstatt and La Tène cultures. It also includes information on the location of Celtic and non-Celtic peoples of the era. Please note that this map is an amalgamation of information included in three separate maps in the Atlas of the Celtic World (Haywood, 2001). With the power of overlaying in transparent format, the composite map (Fig. 6) clearly explains the aforementioned historical evolution of Celtic language.

Another example is the mapping of the shifting locations of folk music festivals in West China conducted by Professor Dwyer at University of Kansas (Fig. 7). The

Fig. 4. A sample procedure to harvest a WMS service

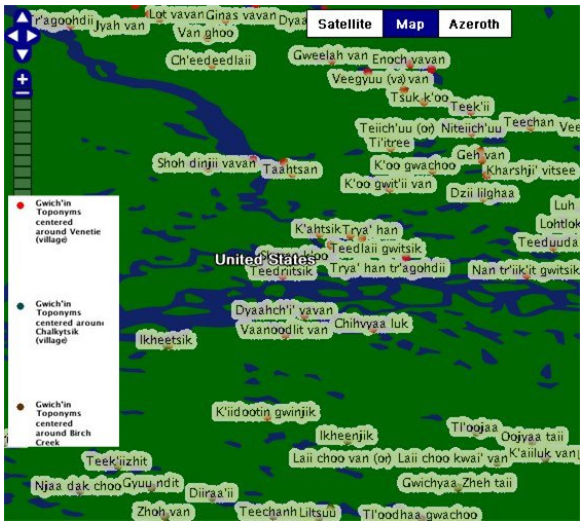

Fig. 5. Gwich'in Toponyms in Alaska

map shows these events happened from April to July in a remote area near the border of Gansu and Qinghai provinces in West China. This area is known to be inhabited by the mixtures of ethnic groups (Tibetan, Mongol and Hui in addition to Han Chinese). These events provided excellent opportunities for the audience to speak their own native languages as a way to preserve their own cultural heritages. A map showing stamped routes in conjunction with the information about landscape and demographic composition helps interpret the significance of the folk music festivals in keeping local culture and language alive.

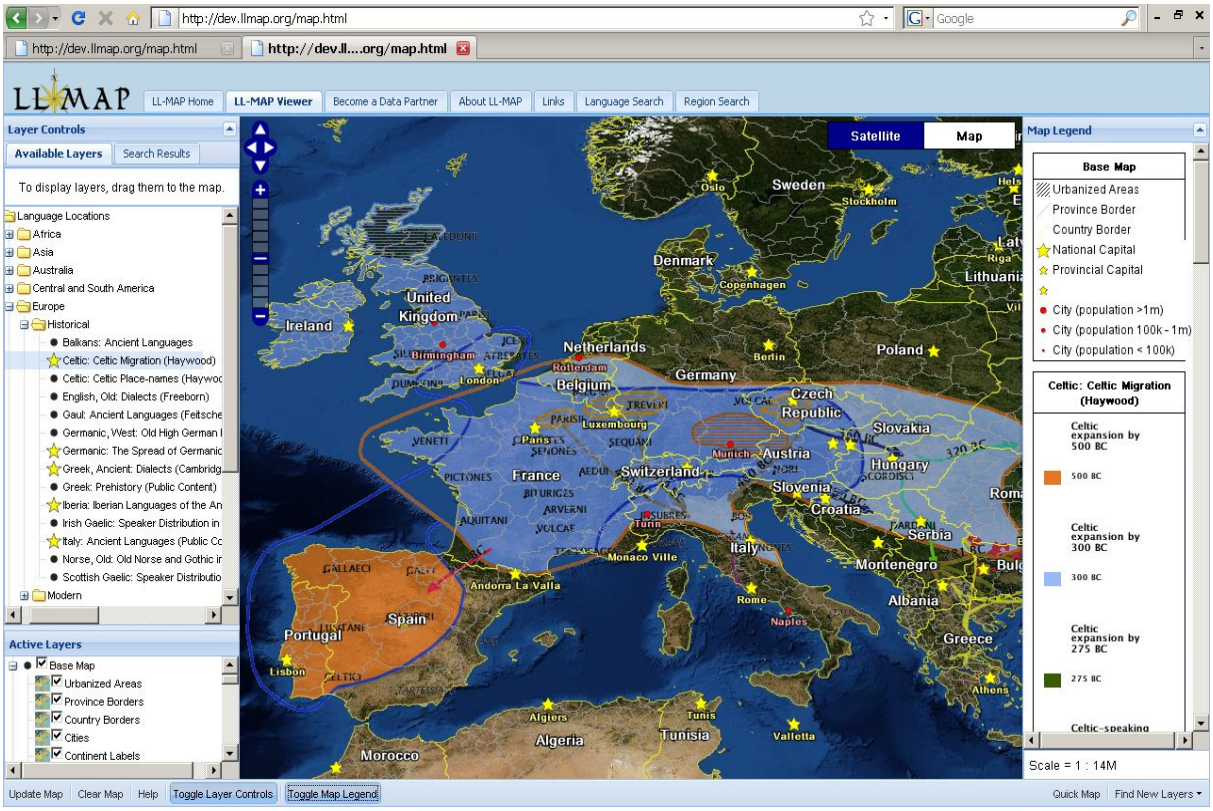

Fig. 6. Celtic Migration

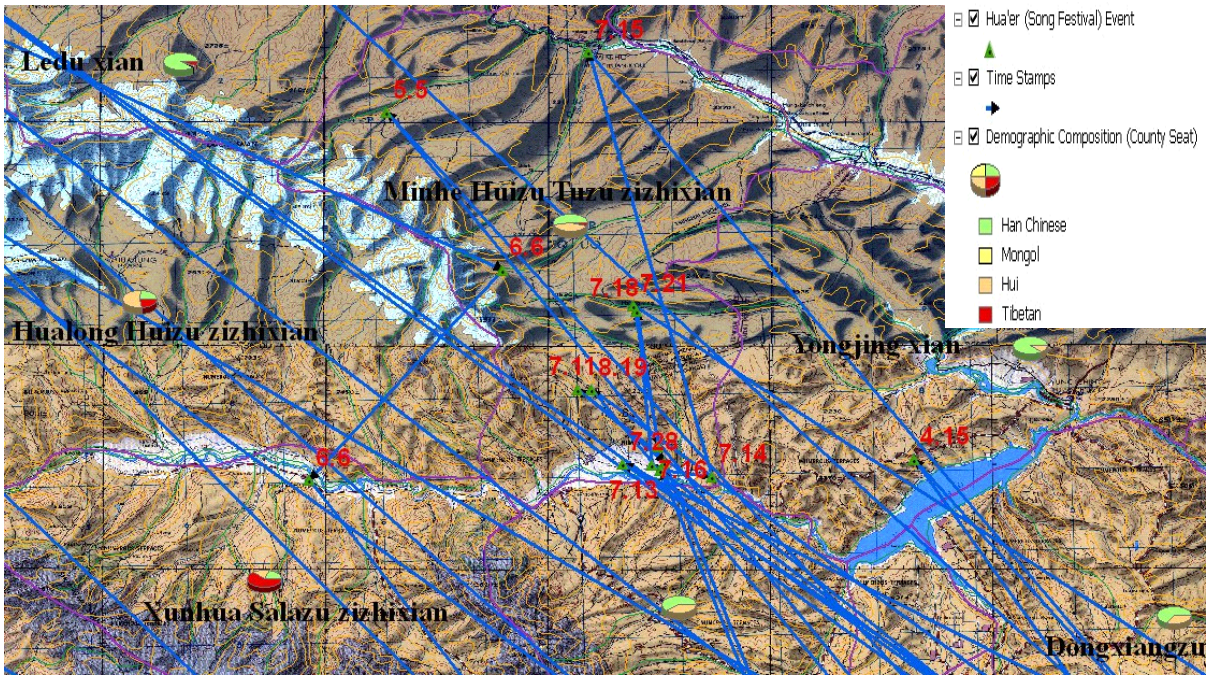

Fig. 7. The shifting locations of folk music festivals in West China.

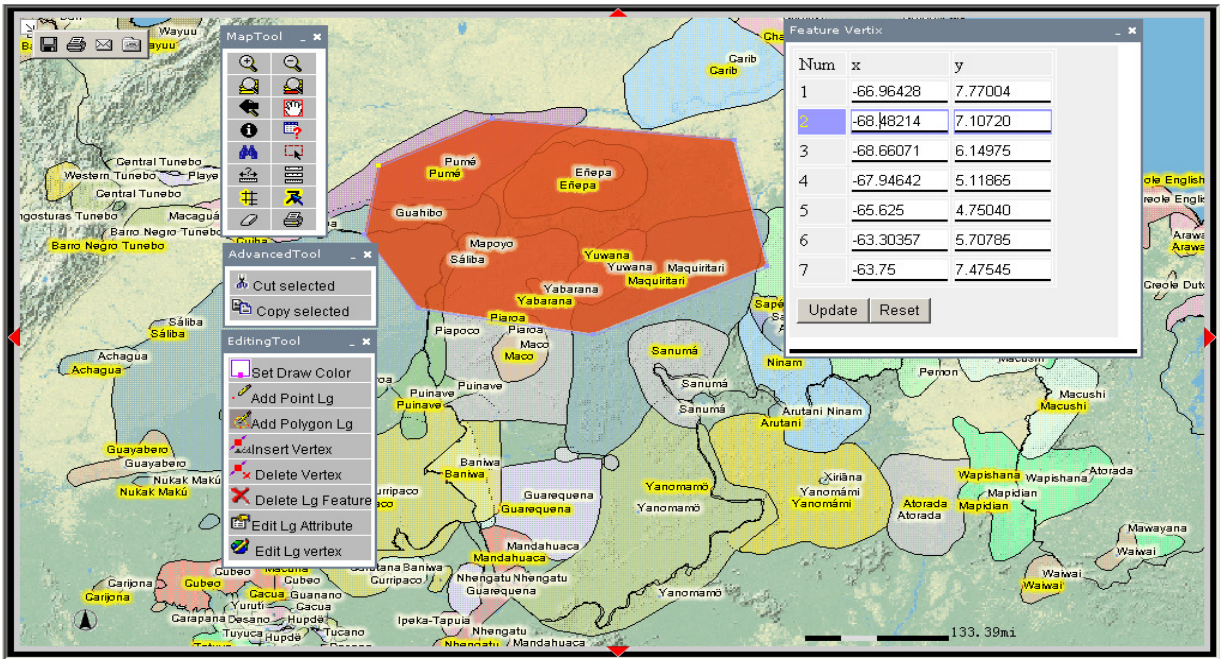

Fig. 8 A polygon showing an area in which a language is spoken is selected and the seven vertices are shown with coordinate values. Linguists can update the shape of the selected feature by Delete Vertex, Insert Vertex or Edit Lg Vertex (updating the coordinate values of vertexes). The attributes of the feature can be added by the tool "Edit Lg Attribute" as shown in the EditingTool Panel.

VII. Collaborative Research Toolbox

Linguists are very interested in seeing tools that will allow them to input their findings and share their findings with others. They would thus like to have a mark-up toolkit which enables them to create and update data or mark their comments spatially. Although there are several online tools that support spatial data editing within a local network, no publicly-accessible online tools are yet available because of the complexity of online data creation. Online data editing or commenting not only involves attribute data (e.g. language name, language speakers) but also spatial data (e.g. where the language is spoken) on a website. Though this is a complex operation, the LLMAP team has developed a flexible on-line spatial editing toolset embedded within the LLMAP portal to handle this issue (Fig. 8). A common framework for spatial data editing is also available on http://geodata.acad.emich.edu/realtime. Thus any linguist can add a new polygon representing where a specific language is spoken or change the

shape of an existing language spoken area by moving the vertex of that polygon shape.

VIII. Conclusion

The LLMAP portal is one of the first comprehensive and operational Web-based linguistic data management systems and analytical toolboxes. LLMAP is designed with three primary goals: 1) a cutting-edge geographically enabled Web-mapping system for linguistics; 2) a portal supporting transparent and interoperable integration of all possible linguistic and related data and sources; and 3) a platform to support collaborative research in linguistics.

The integrated geo-databases embodied in the LLMAP system support innovative ways to look into linguistics data in relationships to environment, landscape, demographics and socioeconomics. LLMAP supports indexing and quick search of places and associated language information and also incorporates temporal and spatial information connected by relationships inferred from historical interpretation, and the visualization of these data through linked maps and timelines. Thus it promotes innovative research methods, and these in turn may lead to new insights into the relationships (both historic and current) among languages, physical environments and human populations. The linked views and acetate dynamic map layers are technically advanced data visualization functions. The aforementioned illustrations of toponymical analysis and temporal and spatial analysis are good examples of linguistic research.

The LLMAP portal is transparent and interoperable. Its user interfaces are designed to increase usability, improve computational efficiency, and enhance response time. The main interfaces are ArcGIS-based using Javascript tool kits and libraries, asynchronous requests, and the latest web standards in XHTML.

The creation of Javascript objects, with constructors to represent types of map interfaces, allows for proper function and variable scoping, while still having the ability for the driver program to synchronize the interfaces. The LLMAP system is thus made more integrative, extensible, transparent, and easy-to-use for non professionals. More importantly, the Web data harvesting toolkit enable the LLMAP system to fetch other public online sources and to generate new maps in real time, which makes the system open and expandable.

The nature of collaborative linguistic research requires distributed data access and analysis (preferably synchronized), and real-time group-decision support toolsets, such as expert make-up tools (editing and drawing), discussion forums, and video-Skyping. The LLMAP system has made a significant effort in supporting real-time editing and sketching.

However, due to the limitation of development time and the long process of integrating or incorporating data from partnering organizations, some toolkits demonstrated so far are not fully interoperable at this time. The LLMAP system is still in the fine tuning and debugging stages, though it is almost fully functional online. Furthermore, the online technology as well as geospatial technology is advancing quickly, so it is still a challenge to keep the LLMAP system current and up to new standards.

IX. Acknowledgement

The LLMAP project is partially supported by a grant from US Natural Science Foundation (Award ID: 0527512). The authors want to thank the US National Science Foundation for its generous support, but will take full responsibility for what has been expressed in the manuscript.

REFERENCES.

[1]. Goodchild, M. F. and Hill, L. L. (2008). *Introduction to digital gazetteer research.* International Journal of Geographical Information Science , 22, 1039-1044.

[2]. Gordon, Raymond G., Jr. (ed.) (2005). *Ethnologue: Languages of the World,* Fifteenth Edition . SIL International, Dallas, Texas (Online version: http://www.ethnologue.com/).

[3]. Guo, Q., Liu, Y., and Wieczorek, J. (2008). *Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach.* International Journal of Geographical Information Science , 22, 1067-1090.

[4]. Hastings, J. T. (2008). Automated conflation of digital gazetteer data. International Journal of Geographical Information Science , 22, 1109-1127.

[5]. Haywood, J. (2001). *Atlas of the Celtic World,* Thames & Hudson, London.

[6]. Janowicz, K., and Kebler, C. (2008). *The role of ontology in improving gazetteer interaction.* International Journal of Geographical Information Science , 22, 1129-1157.

[7]. Jones, C. B., Purves, R. S., Clough, P. D. and Joho , H. (2008). *Modelling vague places with knowledge from the Web.* International Journal of Geographical Information Science , 22, 1045-1065.

[8]. Veselinova, L.N. and Booza, J. C. (2009). *Studying the multilingual city: a GIS-based approach.* Journal of Multilingual and Multicultural Development , 30, 145-165.

[9]. Young, J. A. (2004). *Harvesting and Resolution Methods for Building OAI-based Services,* CERN OAI3 Workshop # 4, Geneva, Switzerland, 2004 (http://eprints.rclis.org/1011/1/tutorial4young.pdf)