

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Issue Statement

Ljuba Veselinova
Stockholm University

In the statement below I focus on the output of commonly used map servers and how it challenges interoperability. Laurini (1998) makes a distinction between two kinds of interoperability: (i) for a program, data interoperability means the ability to utilize a range of data formats. (ii) for a dataset, program interoperability means that it can be used by different types of programs and subsequently for a variety of purposes. In the discussion below I concentrate on the second aspect.

In recent years, Geographical Information Systems (GIS) have become more available and more user friendly. As such they are no longer confined to geography departments and city planning but have become more and more pervasive in all kinds of domains, including the humanities. Slowly but surely linguists are discovering their potential as a powerful research tool in various aspects of language studies (see (Haspelmath et al., 2005) for an example of the use of GIS in linguistic typology). Currently, there are a number of linguistic projects where one of the main goals is to view languages in their proper geophysical context. This in turn means that a geographical component and subsequently geo-(linguistic) data are being added to current linguistic databases. Most geographically oriented linguistic projects are currently in their initial stages. While very diverse in coverage and scope, it seems that the emerging tendency is to serve the available geolinguistic data on the web by means of map servers. Several commercial desktop application dominate the geographical information industry: ESRI ArcInfo (webserver ArcIMS), SmallWorld GIS (webserver Internet Application Server), Intergraph Geomedia (webserver Geomedia WebMap), MapInfo Professional (webserver MapXtreme) (see (Zhang et al., 2003) for a more detailed inventory of existing software products). The issues for interoperability usually associated with proprietary software apply to the products mentioned above as well: incompatible data models, database storage structures and file formats. In addition, the current praxis is to use internet GIS programs to deliver spatial data through transmission of raster images (GIF or JPEG formats) over the web. I consider the latter transfer a serious problem for interoperability. Specifically, by transferring images only, actual data sharing does not occur since image transfer does not allow the selection of specific features. Besides, image transfer is also resolution and platform dependent.



The Open Geospatial Consortium has created Geographic Markup Language (GML) to facilitate data exchange. GML is XML grammar written in XML schema for the modeling, transport and storage of geographic information including both the spatial and the non-spatial properties of geographical features. The linguistic community should gain awareness of GML and GML based tools for the exchange of spatial data in linguistic databases. GML based servers can deliver vector data by styling the data into scalable vector graphics (SVG). Zhang et al (2003) cite several advantages for delivering SVG vector data rather than raster GIS data: (i) Compatibility: SVG uses text based XML format which is compatible with other formats. (ii) Graphic quality: SVG format graphics are scalable and resolution independent. SVG data can be scaled without loss of quality across platforms and devices. (iii) GML based databases can let users exchange data on feature level (iv) GML allows for more flexibility in general. It only defines a basic geographic feature schema and a geometry schema. Based on these schemas, users can define their own specific schemas for their own spatial data documents. (iv) Users can develop their own schemas according to GML 3.x. Thus the need for standardization is reconciled with the need for diversity by providing standard means of extending the GML format.

Since the use of GIS and map servers in linguistics is still very much in its beginnings, it appears that it would be especially suitable to discuss some best practices and recommendations such as specify and motivate preferences for commercial versus open source software, whether we should deliver features rather than images over the web, and finally a linguistic extension to current GML.

References

- Haspelmath, Martin, Dryer, Matthew, Gil, David, and Comrie, Bernard eds. 2005. World Atlas of Language Structures. Oxford: Oxford University Press.
- Laurini, R. 1998. Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. *International Journal of Geographical Information Science* 12:373-402.
- Zhang, Chuanrong, Peng, Zhong-Ren, Li, Weidong, and Day, Michael J. 2003. GML-Based Interoperable Geographical Databases. Ms.

