Preparing the Digital/GIS Language Atlas of China for the ECAI e-Publication Program

Lawrence W. Crissman
Director, Spatial Data Projects
Griffith Asia Pacific Research Institute (GAPRI)
Griffith University, Brisbane, Australia

The Language Atlas of China (LAC) was published by the Longman Group (Far East) in two parts, both dated 1987, the first of which contained only somewhat less than half of the maps included in the second, complete 'Part II'. This is worth mentioning because some libraries seem to be unaware that they hold only the first, incomplete version. The Australian Academy of the Humanities sponsored the work, which was a joint effort involving the Chinese Academy of Social Sciences, under the general editorship of Professor August Wurm, Australian National University. The maps were prepared by Theo Bauman, a cartographer in the Department of Linguistics at ANU. The linguistic scholarship on which the maps were based, which is presented in the text pages of the Atlas, was assembled by various persons at the CASS and translated and compiled into a standardised format for publication by Mei W. Lee.

Soon after first obtaining Australian Research Council funding in 1992 to establish the Spatial Information Infrastructure for Asian Studies in Australia (SIIASA), with the encouragement of Professor Wurm I obtained copyright permission to produce a GIS version of the LAC from the Australian Academy of the Humanities and the Longman Group (Far East). With the cooperation of Theo Bauman I obtained a loan of the colour separates used in printing the LAC. They could be scanned and automatically vectorised with the technology then available, which worked well with black and white line work but not with coloured maps such as those in the published atlas . However, as latitude/longitude were not published in the same colour as the lines dividing the polygons for the various languages and dialects shown on the maps, the vector data had to be rubber sheeted in a labour intensive process that was not free of interpretive procedures that are still causing problems in finalising the GIS versions of the maps.

The scanning and vectorisation of the colour separates to produce MicroStation .dgn files was done in 1993, but other priorities sidetracked further work until 1998 and 1999 when time and resources allowed a more or less coherent set of GIS data derived from the four general 'A Series' maps covering all of China. At that time, vector data for most of the other maps was also brought into the ESRI ARC/INFO 7 environment, and MapInfo versions of the individual map sheets were produced. Then, as the ECAI e-Publication Program began to take shape and it was proposed that the Digital/GIS Language Atlas of China be included, further work was undertaken during the first half of 2001 using ArcView 3.2 to classify and attribute the language supergroups, groups, and dialects, etc., represented on the map sheets. Finally, in early 2002 the GIS data were brought into the new ArcGIS 8 environment where a Geodatabase was built that combines the data for particular languages contained on different map sheets in the LAC.

During the second half of 2001, when I was at McGill University on a sabbatical, I was able to arrange for the maps in the LAC to be scanned at high resolution by the Rare Book section of the McGill Library to produce the images of the maps that will be included in the e-Publication. Using equipment at the Centre for East Asian Research in the Department of East Asian Studies, I was also able to scan the text pages of the LAC and apply optical character recognition software to produce word document versions of them. The IPA glyphs and Chinese characters have been added in UTF-8 at Griffith University since my return in the beginning of 2002. The final HTML versions of the text pages will have the tables in the text pages inserted as .jpg files in order to avoid formatting problems and allow the inclusion of the symbols for tones used in the LAC which are not contained in any readily accessible font.

The Digital/GIS Language Atlas of China being produced for the ECAI e-Publication Program will therefore have a number of components, including a detailed technical discussion and background information on the histories and linguistic classifications of the Minority Nationalities of China and government policies towards them. One section will contain the scans of the map sheets, which will be geo-registered so that vector data can be overlain on them. Another will contain the HTML documents that reproduce the text pages in the LAC. The GIS spatial data produced from the colour separates will be included in two forms, one matching the separate map sheets, and another that combines all of the information on particular languages and dialects from various sheets of the maps apart from the small scale, generalised A1 to A4 sheets that cover all of China. County boundary data matching the date of the LAC will also be included so that they can be overlaid on the map scans and the GIS data.

The Digital/GIS Language Atlas of China, as published by ECAI, will provide users with several advantages over the paper version. First, the latter is not always easy to find, as not all university libraries hold a copy. Those that do have often put theirs into their Rare Book or other restricted collections, which can sometimes only be accessed with difficulty even by researchers at those institutions. Second, the scale and projection parameters of the published maps is not indicated, and latter is only known through personal communication from the cartographer. Third, the language regions displayed on the paper maps in the LAC are difficult or impossible to accurately associate with other kinds of information, such as topography or local administrative units. Only provincial-level administrative boundaries are depicted, along with a few selected cities and larger towns, while only main rivers and their major tributaries are included. However, when the information on the maps is reproduced in vector GIS format, in decimal degrees or in a known projection/scale, the resulting spatial data can be use in conjunction with other documented GIS resources to overlay the language and dialect distributions with other kinds of data, such as topography and various demographic and socio-economic statistics. That will greatly enhance the utility of the information so painstakingly assembled and presented in the original Language Atlas of China.