Considerations for Harmonising Cross-Border Geospatial Datasets

Gobe Hobona[1], Carmelo Attardo[2], Robert Laurini[3], Mike Jackson[1], Maria Pla[4], Stefania de Zorzi[5], Anette Breu[6], Catherine Roussey[2], Alina Kmiecik[7]

[1]Centre for Geospatial Science at the University of Nottingham, UK
[2]Intergraph Corporation, Italy
[3]Institut National des Sciences Applique'es de Lyon, France
[4] Institut Cartogràfic de Catalunya, Catalonia, Spain
[5]Corila, Venice, Italy
[6] Intergraph Corporation, Germany
[7] Intergraph Corporation, Poland

EXTENDED ABSTRACT

INTRODUCTION

The publication of the INSPIRE directive has provided a legal framework through which a European Spatial Data Infrastructure (ESDI) can be established (EC, 2007-2009). A key aspect of establishing an ESDI will be in providing guidelines and mechanisms for harmonising various datasets from the 27 member states of the European Union. The European Commission has led the development of implementation rules to help member states adopt the INSPIRE directive. Some of the INSPIRE implementation rules involve the development of common data models for themes listed in Annexes of the directive. Once the final versions of the common data models have been published, it will then be necessary for member states to harmonise their existing datasets into the common data models. Currently, INSPIRE Thematic Working Groups (TWGs) have successfully developed draft versions of data models for Annex I themes.

In parallel, the GIS4EU project was commissioned to develop guidelines and tools for adopting common data models based on specific themes in Annex I and Annex II of the directive (GIS4EU, 2009). The Annex I themes selected were administrative units, transport networks and hydrography. The Annex II theme selected was elevation. To avoid duplicating effort, GIS4EU does not intend to develop new common data models for Annex I themes but instead has provided a critical analysis of the draft INSPIRE data models in relation to datasets supplied by GIS4EU data providers. However, as the INSPIRE Annex II data models have not been developed yet, GIS4EU has successfully developed a common data model for the Elevation theme. This extended abstract presents some of the considerations on data harmonisation identified by the GIS4EU project. The considerations are described at a high level to ensure applicability within a variety of computing platforms. The workshop presentation will recommend how to address each considered scenario. The rest of this paper presents a classification of the considerations according to scale, geometry and the attributive nature of the data being integrated.

MERGING UNIFORM SCALE CROSS-BORDER DATASETS

This section presents considerations related to the merging of cross-border datasets with similar scale. The rules presume that a target schema is available, such as those from the INSPIRE TWGs. The rules can be categorised into those that affect *geometry* and those that affect *alphanumeric attributes*. The following is a listing of some of the rules affecting geometry, additional rules will be presented during the workshop, with illustrations of the situations for application of the rules:

- Integration of input datasets in different coordinate reference systems requires the datasets to be reverse projected into the ETRS 1989 coordinate reference system, which is recognised as the European standard in INSPIRE implementation rules.

- If input datasets are in different coordinate reference systems or have been surveyed by different individuals, then it is possible that the geometries may overlap or there may be gaps been geometries along the borders. In such a scenario, boundary force fitting may be applied by moving adjacent boundaries to the midline defined by the Hausdorff distance between the boundaries (Hangouet, 1995). However, any boundary force fitting will make the datasets inapplicable within any legal context.

- If integrating raster datasets with different cell geometries or orientations, gaps or overlaps may occur between the datasets. It may be necessary to fill the gaps between the datasets. However, if the gaps are larger than the cell sizes of both datasets then filling the gaps results in a significant introduction of incorrect data. Therefore, if the gaps between two raster datasets are larger than both their cell sizes then

the gaps should be left 'blank' with a 'No value' entry which is supported by most GIS. If the gap is smaller than the cell sizes of the datasets, the datasets can be converted into a triangular irregular network (TIN), after which an interpolation can be made between adjacent cells.

The following is a listing of some of the rules affecting alphanumeric attributes, additional rules will be presented during the workshop, with illustrations of the situations for application of the rules:

- The attributes of the input datasets should be mapped to attributes in the target schema. In the case of thematic INSPIRE data specifications the attribute mapping will effect a language translation into English. However, attribute values are considered data; therefore if there is no definite English equivalent then the attribute values should not be modified. Translation of placenames should be avoided as most placenames do not have an English language equivalent. We have noticed that a semantic reference system (Kuhn, 2005) agreed on by all INSPIRE member states does not currently exist. In several cases the feature types proposed in the INSPIRE data specifications require additional detail to explicitly define the level of detail stored in each national dataset. This leads to a partial translation between schemas and consequently some information is lost.

- Considering feature types and attributes from different datasets to be from different national semantic reference system or national ontologies, semantic similarity (Rodríguez and Egenhofer, 2003) approaches could be applied in the mapping of multilingual attributes.

- When integrating raster datasets, it is necessary to reclassify pixel values into a uniform classification. If the pixel values show elevation it is possible that the values may be in different units for example Metres and Centimetres. It is then necessary to apply a conversion factor through raster algebra. Conversion factors can also be applied to attributes in vector datasets, for example, when integrating transport data with distance shown as Miles with another showing distance in Kilometres.

MERGING MULTIPLE SCALE CROSS-BORDER DATASETS

This section discusses the considerations related to the merging of datasets at different scales and rules for handling the degradation of quality and resolution. The considerations are primarily applicable to the merging of datasets from local authorities covering different levels of jurisdictions for example, a city in a district and a district in a province. Most of the rules are inherited from cartographic generalisation studies because of the degradation of data quality of the large scale dataset in order to merge it with the smaller-scale dataset. Some of the considerations are:

- If the datasets are of significantly different scales then the merged dataset may appear clattered. To improve clarity, some of the least relevant feature types should be selected and removed. Relevance can be judged based on theme, for example in a transport network dataset, rivers are more relevant than buildings (e.g. in Venice, Italy), whereas rivers are less relevant than roads (e.g. in Turin, Italy).

- Often when merging datasets of significantly different scales, it is necessary to amalgamate some of the features, for example a big area covered by a set of small lakes may be represented by some aggregations of some of the smaller lakes.

- For area or linear features, the geometries can be collapsed. For example, a river could be represented as a line along the centreline rather than as the polygon defined by the river bank, or a road service area could be represented as a point rather than as an area.

- There are times when the difference between the scales of datasets is small. A generalisation operation that is suitable for this scenario is simplification of geometries. This involves reducing the vertices or coordinates within a geometry resulting in a 'smoother' feature. However, care should be taken that the simplified feature maintains topological consistency in relation to nearby features. For example, a simplified road should not introduce overlaps with an adjacent river.

There are several other generalisation operations that could be applied, a summary of them can be found in Longley et al (2005). The workshop presentation will present only those that are applicable to GIS4EU.

CONCLUSIONS AND FUTURE WORK

This abstract has presented some of the considerations applicable during harmonisation of different datasets from adjacent countries or different levels of jurisdiction in the same county. Although some of the considerations are built-into current commercial-of-the-shelf GIS, other considerations depend on a user making an informed decision on the appropriate rule to apply to a given data integration scenario; it is envisioned that this abstract will assist organisations in making such an informed decision. Future work within GIS4EU will involve aggregation of supplied datasets using Intergraph Geomedia Fusion software and also development of a geoportal to assist the data harmonisation process.

ACKNOWLEDGEMENTS

The GIS4EU project is funded by the European Commission through the e-ContentPlus programme.

BIBLIOGRAPHY

EC. (2007-2009) INSPIRE Directive and Data Models, Last visited 22/03/2009, Available at: http://inspire.jrc.ec.europa.eu

GIS4EU (2009) GIS4EU Project Website, Last visited 22/03/2009, Available at: http://www.gis4eu.eu

OGC, (1999-2009) Open Geospatial Consortium specifications, Last visited 22/03/2009, http://www.opengeospatial.org

HANGOUET, J.F. (1995) Computation of the Hausdorff distance between plane vector polylines, Twelfth International Symposium on Computer-Assisted Cartography, Charlotte, North Carolina

KUHN, W. (2005) Geospatial Semantics: Why, of What, and How? Journal on Data Semantics (Special Issue on Semantic-based Geographical Information Systems, Spring 2005, LNCS 3534): 1-24

LONGLEY P. A., GOODCHILD M. F., MAGUIRE D. J., RHIND D. W. (2005) Geographic Information Systems and Science, West Sussex England: Wiley

RODRÍGUEZ, A. AND EGENHOFER, M. J. (2003) 'Determining Semantic Similarity among Entity Classes from Different Ontologies', IEEE Transactions on Knowledge and Data Engineering, 15, (2), pp. 442 - 456