Automating Standards-Based Metadata Creation using Free and Open Source GIS Tools

Claire Ellul[1], Nart Tamash[1], Feng Xian[1], John Stuiver[2], Patrick Rickles[1]

[1]Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT UK

Tel. +44 (0) 20 7679 4118 Fax +44 (0) 20 7679 3042

c.ellul@ucl.ac.uk

[2]Laboratory of GeoInformation Science and Remote Sensing, University of Wageningen, 6708 PB Wageningen, The Netherlands

KEYWORDS: Metadata Automation, FOSS GIS, Open GIS, Spatial Database, Academic Spatial Data Infrastructure

Summary

The importance of understanding the quality of data used in any GIS operation should not be underestimated. Metadata (data about data) traditionally provides a description of this quality information, but it is frequently deemed as complex to create and maintain. Additionally, it is generally stored separately from the data leading to issues with updates to the data not being reflected in the metadata and to users not being aware that metadata exists. These issues have gained increasing importance as more and more non-specialists access GIS software and data – driven by Free Open Source Software (FOSS) and Open Data, particularly in an academic context. This paper describes an approach to address these two issues – tightly coupling data and metadata and automating many elements of standards-based metadata creation. The tools have been developed using the FOSS packages Quantum GIS and PostGIS.

1. Introduction

Metadata has long been understood as a fundamental component of any Geographical Information System (GIS) data management process, providing information relating to discovery, evaluation and use of datasets and describing their quality. It provides a formal description of the data quality (Kim 1999), allows for data reuse (Craglia et al. 2008) and avoids data duplication. Having good metadata about a dataset is fundamental to using it correctly and to understanding the implications of issues such as missing data or incorrect attribution on the results obtained for any analysis carried out. Traditionally, spatial data and the corresponding metadata was created by expert users (e.g. national mapping agencies). Increasingly, however, data used in spatial analysis comes from multiple sources and could be captured or used by non-expert users – for example academic researchers - many of whom are from non-GIS disciplinary backgrounds, not familiar with metadata and perhaps working in geographically dispersed teams. This greater uptake of GIS is being furthered by the availability of Free and Open Source (FOSS) GIS packages and increasing volumes of open data. To support and further this open data sharing, high quality metadata is required to allow users to discover and evaluate data and use it appropriately.

This paper describes an open-source approach to addressing two outstanding issues in metadata creation and maintenance that are of particular importance when considering non-

expert users such as academics. Firstly, the creation of metadata is a tedious and costly task which is often left until the end of a project and completed to the minimum standard possible. This results in metadata that is barely useful and often contains errors (West and Hess, 2002). Secondly the separate storage of metadata and data (as shown in in Figure 1) means that metadata is not automatically updated when the data changes. Tools to create metadata are often separate from the main GIS package.

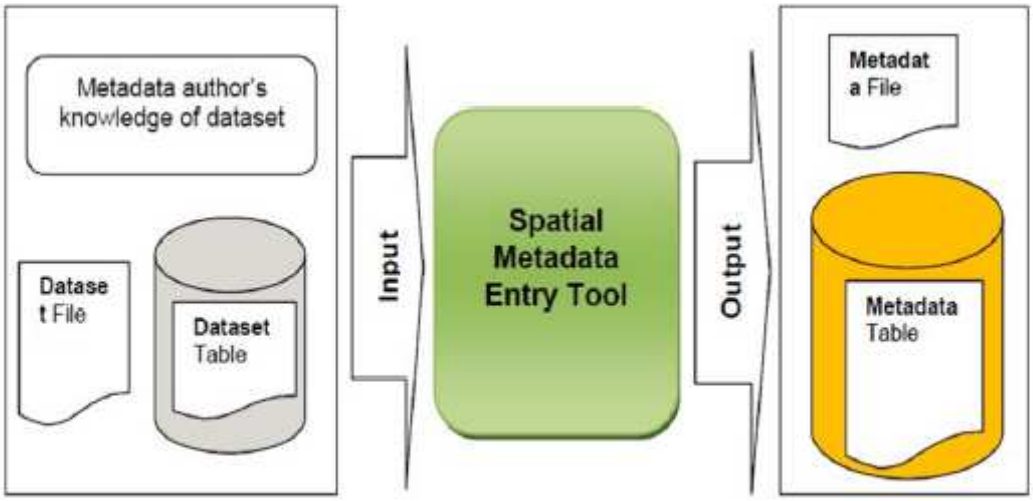

Figure 1. Workflow of metadata creation, update and Storage (adapted from Olfat et al, 2012)

2. Potential for Automatic Metadata Creation

Metadata was derived from 'marginalia', where information such as the map legend, date of production, map projection and producer were printed in the margins of a paper map (Goodchild 2007, Poore and Wolfe 2010), and subsequently influenced by library standards (Poore and Wolfe 2010). To enable interchange and understanding by computer-based systems, metadata is often stored in a very structured, standardized format (for examples see the United States Federal Geographic Data Committee[1], the Dublin Core Metadata Initiative[2], the International Standards Organization's 19115:2003 Geographic Information Metadata Standard[3]). A 2005 study (Moellering, 2005) identified 22 standards still in wide use and hundreds of smaller standards in limited use. Of particular relevance to users in the European Union is INSPIRE (INfrastructure for Spatial InfoRmation in Europe, INSPIRE 2011a, INSPIRE 2011b) standard. This has been used in this project.

Given the intricacy of metadata standards, even for specialists the complexity of creating and maintaining such metadata is considered significant (Poore and Woolf 2010, Manso-Callejo *et. al* 2010, Batcheller 2008, Craglia et al. 2008). However, a review of the INSPIRE metadata standard reveals that the population of a significant number of elements can be automated when the standard is applied in an academic context, as shown in Tables 1 and 2 below. In particular, it may be possible to automate the population of all of the mandatory elements of the standard. Overall, of the 38 elements listed, 23 have been implemented in the prototype described in the next section.

[1] http://www.fgdc.gov/metadata
[2] http://dublincore.org/
[3] http://*portal.opengeospatial.org/files/?artifact_id=6495*

Table 1. Key to Automated Metadata Elements

Colour Code	Interpretation	Number of Elements
	Implemented – manual creation	3
	Implemented – automatic creation	20 (17 mandatory, 3 optional)
	Not implemented – automation may be possible (perhaps with manual verification)	15 (5 mandatory, 10 optional)
	TOTAL	**38**

Table 2. Automated Metadata Elements

Elements Title	Manual/Automatic Population
1. Resource title	Inserted manually. If not inserted by the user, default value is as dataset name.
2. Resource abstract	Inserted manually.
3. Resource type	Default to 'dataset'.
4. Resource locator	To be implemented if uploaded datasets will automatically be exported to shapefiles and zipped on a server – in this case take the zipped file location (OPTIONAL)
5.1 Identifier code	Take the Object Identifier of the PostgreSQL database
5.2 Identifier namespace	Default to 'ucl.ac.uk_CEGE_metadata_projectname'
6. Resource language	To be implemented using the language identification algorithm currently applied on 'metadata language'. (OPTIONAL)
7. Topic category	List provided by INSPIRE. It may be possible to compare to keywords and automatically assign from a lookup table.
8. Keyword(s)	To be implemented by 'concatenating' all text fields of the dataset and picking the top repeating ones.
9.1 Vocabulary title	Keyword originating vocabulary. Lookup tables possible (lists from INSPIRE or created by the end users). Values would be assigned using the same implementation as 'topic category'. (OPTIONAL)
9.2 Vocabulary reference date	Implementation based on *Vocabulary Title* choice (OPTIONAL)
9.3 Vocabulary reference date type	Code list: creation, last revision, publication. Implementation based on selected *Vocabulary Title.* (OPTIONAL)
10.1 BB northbound lat.	PostGIS function, coordinates in WGS84.
10.2 BB eastbound long.	PostGIS function, coordinates in WGS84.
10.3 BB southbound lat.	PostGIS function, coordinates in WGS84.
10.4 BB westbound long.	PostGIS function, coordinates in WGS84.
11.1 Temporal extent (start)	Maybe implementable by looking for 'date' type columns/fields in the dataset. (OPTIONAL)
11.2 Temporal extent (end)	Maybe implementable by looking for 'date' type columns/fields in the dataset. (OPTIONAL)
12. Date of publication	Default to the date that data was uploaded to the system. Manual verification required (OPTIONAL)
13. Date of last revision	Default to the date the data was uploaded to the system. Update any time data edited. (OPTIONAL)
14. Date of creation	Default to the date the data was uploaded. Manual verification required (OPTIONAL)
15. Lineage	Inserted manually. Might be possible to populate this field downstream by logging the user's activities.
16.1 Resolution scale	Might be possible by comparing to boundary reference data. (OPTIONAL).
16.2 Resolution distance	Might be possible by comparing to boundary reference data. (OPTIONAL).
16.3 Resolution measure unit	Might be possible by comparing to boundary reference data. (OPTIONAL).
17. Conformity degree (the degree to which the dataset conforms to a	Default to false as data will most likely not be compliant to INSPIRE data specifications

specific INSPIRE specification document)	
18.1 Conformity specifications	Default is NULL (assume that conformity is false)
18.2 Conformity specifications date	Default is NULL (assume that conformity is false)
18.3 Conformity specifications date type	Default is NULL (assume that conformity is false)
19. Limitations on public access	Code list available from INSPIRE, assign default value based on this list.
20. Conditions of use	Code list available from INSPIRE, assign default value based on this list.
21.1 Responsible party name	Based on user groups (from database user login) and lookup tables created for these.
21.2 Responsible party email	Based on user groups (from database user login) and lookup tables created for these.
22. Responsible party role	Code list available, it may be possible to assign a default value based on it or create lookup tables when different roles available per user groups.
23.1 Metadata contact name	Based on database user groups and lookup tables created for these.
23.2 Metadata contact email	Based on database user groups and lookup tables created for these.
24. Metadata date	Date when data added to the database. Updated to last date metadata changed if data subsequently edited.
25. Metadata language	Language detection algorithm is used on 'resource abstract' field.

3. Developing the Metadata Automation Software

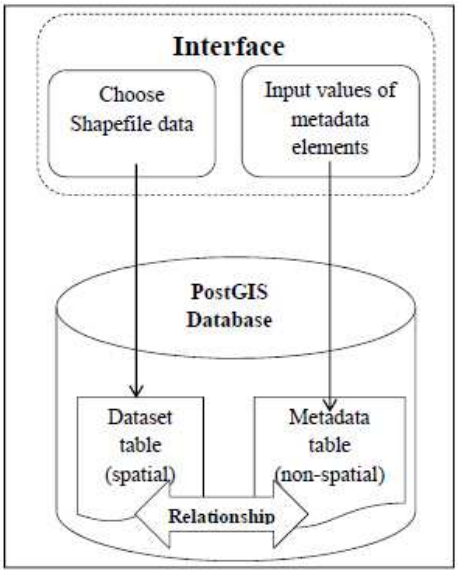

Figure 2. Conceptual Model of the Integrated Metadata System (from Xian 2012)

To enable close coupling of the metadata and data, provide GIS users with easy access to metadata as part of their workflow and ensure automated metadata update, a spatial database forms the back-end of the tools developed, coupled with a GIS front-end. Specifically, two Free and Open Source (FOSS) tools were selected – PostgreSQL/PostGIS[4] and Quantum GIS[5] - to maximise the reach of the resulting software. This is particularly important in an academic context, where non-specialist GIS users may not have easy access to commercial

[4] PostgreSQL 9.4, PostGIS 2.0
[5] Quantum GIS 1.7.4

GIS packages on their own desktops (although these may be available in student labs).

Figure 2 shows an architecture diagram for the tools. The dataset is stored in the spatial database, and is tightly coupled, by means of triggers (processes that run when a dataset stored as a table in the database is created, updated or deleted) to the corresponding record in the metadata table. The overall data and metadata flow is as follows:

1. The user loads the dataset into the database via a custom QGIS plug-in, shown in Figure 3, which permits the user to enter the mandatory metadata elements that cannot be automated - i.e. title, abstract and lineage.

2. On loading, a new metadata record is automatically created, and populated with details about the dataset as per Table 2 above.

3. Once the dataset is loaded, a second custom plug-in allows the user to connect to the database and view the available metadata for each dataset.

4. Should the dataset change in any way (due to user edits), further trigger processes will run automatically to update the metadata record.

As part of the metadata creation process a bounding box for each of the datasets is created, which allows the user to add a 'metadata' layer to the map and provides a quick overview of the locations of available datasets. Clicking on each polygon using the 'information' tool shows the metadata for any datasets at available for the selected, and the metadata can be searched using standard GIS tools, thus providing an inbuilt discovery function (Figure 3 below).

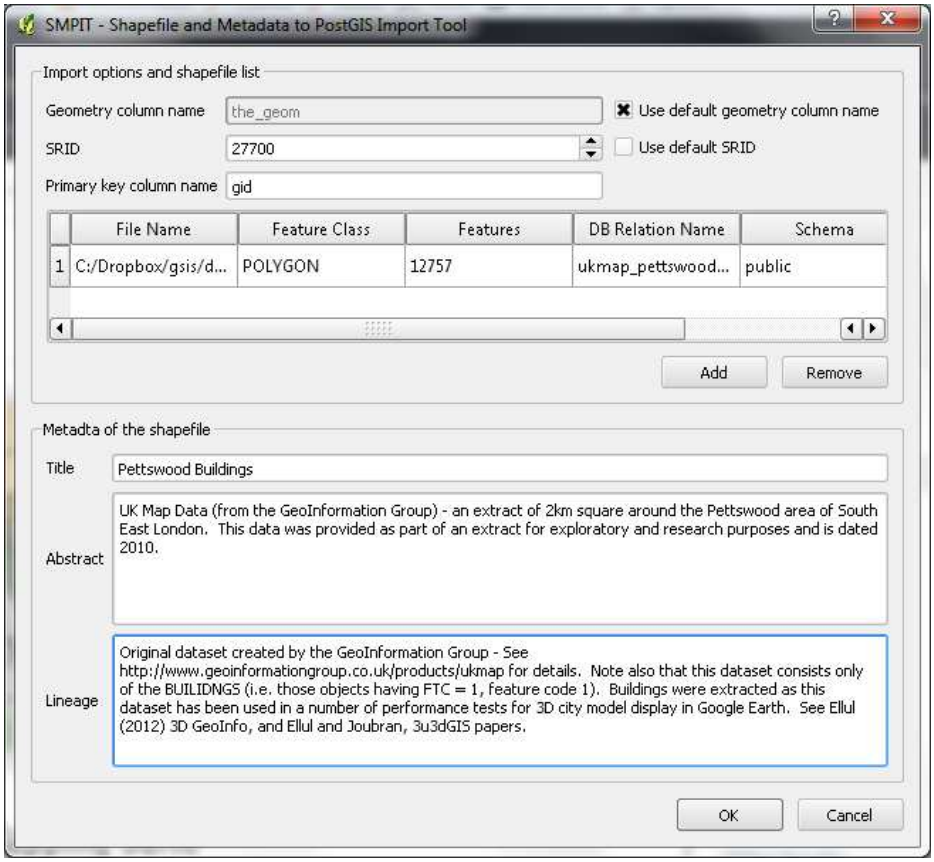

Figure 3. Entering Metadata on Data Upload to PostGIS via QGIS

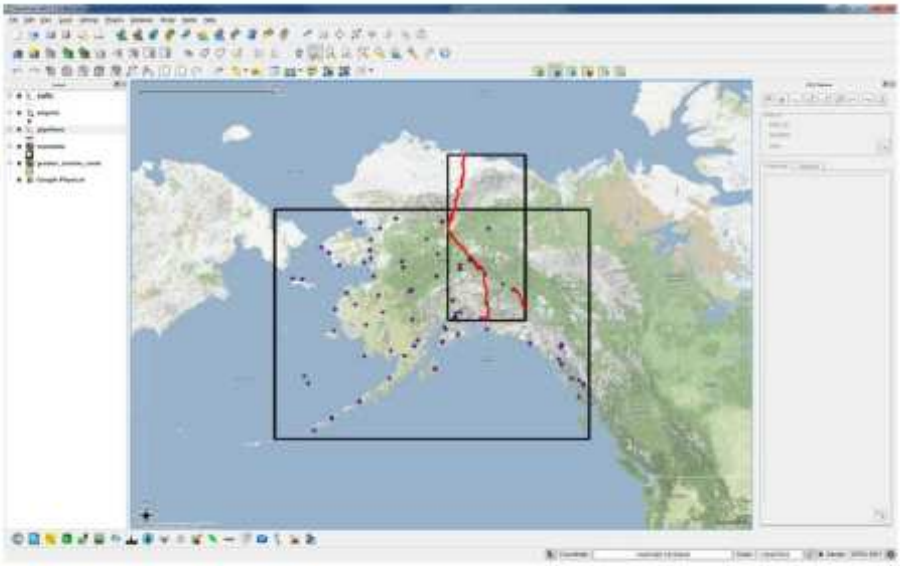

Figure 3. The Resulting "Metadata" Map

4 Discussion and Further Work

The tools and processes described above demonstrate the potential of the automatic creation of many elements of standards-based metadata using FOSS GIS software packages. Importantly, making use of a spatial database ensures that, by the use of triggers, should the data change the metadata will automatically be updated. The two plug-ins developed in QGIS highlight the process of integrating metadata with the standard GIS workflow – making it more difficult to ignore by end users and ensuring that metadata is populated as soon as data is modified. The use of language detection software to automatically populate dataset and metadata language elements is of particular interest, as is the potential for finding keywords by examining the text in the dataset itself.

Storing data in a database offers immediate advantages in terms of sharing data with other users, central backup and multiple levels of user privileges across the data. However, setting up a shared spatial database does require additional technical skills perhaps not within the purview of non-specialist users of GIS. It is envisaged, therefore, that the database could be set up by specialist data managers that are being increasingly employed by universities in order to maximise reuse of research data. Once setup, such repositories could be made available to end users as required.

Further work includes taking the prototype plug-ins and developing a more professional interface for these, as well as integrating the metadata viewer with the process of opening a dataset in QGIS and implementing the code for the remaining metadata elements. Once this is achieved, it is planned to open the source code to the public. As a standards-based spatial database underpins this work, and much of the metadata creation is carried out inside the database, it is also feasible to develop metadata management tools in other software such as ArcGIS.

In the longer term, the standards-based metadata will be extended to cover non-standard issues such as those described in Ellul et al. (2011), which relate to the limitations of standards-based metadata to describe data adequately for end users. This will permit users

to, for example, comment on or rate both the data and the metadata, describe how they have used the data and permit the inclusion of additional project-specific fields to facilitate metadata and data management in an academic context.

References

Batcheller, J (2008). Automating geospatial metadata generation – An integrated data management and documentation approach, Computers & Geosciences, 34, 287-398

Craglia M, Goodchild, M, Annoni A, Camara G, Gould M, Kuhn W, Masser D, Maguire D, Liang S, Parsons E (2008). Next Generation Digital Earth. A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science *International Journal of Spatial Data Infrastructures Research*, 3:146-167

Ellul C, Winer D, Mooney J, Foord J, (2011), Bridging the Gap between Traditional Metadata and the Requirements of an Academic SDI for Interdisciplinary Research in Rajabifard A and Coleman D (eds.) *Spatially Enabling Government,. Industry and Citizens*, GSDI Association Press

Goodchild M.(2007) Beyond Metadata: Towards User-Centric Description of Data Quality. Proceedings of the *International Symposium on Spatial Data Quality*, June 13-15, Enschede, Netherlands

INSPIRE (2011a). About Inspire [online] Available from: http://inspire.jrc.ec.europa.eu/index.cfm/pageid/48 [Accessed 23rd March 2011]

INSPIRE (2011b). INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119 [online] Available from: http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf, [Accessed 23rd March 2011]

Kim T (1999) Metadata for geo-spatial data sharing: a comparative analysis. *The Annals of Regional Science* v33. 171-181.

Manso-Callejo M.A, Wachowicz M, Bernabé-Poveda A (2009). Automatic Metadata Creation for Supporting Interoperability Levels of Spatial Data Infrastructures. In: GSDI 11 World Conference: Spatial Data Infrastructure Convergence: *Building SDI Bridges to Address Global Challenges*, Rotterdam, The Netherlands, June 15-19, 2009. - Rotterdam : 2009

Moellering H. 2005. *World Spatial Metadata Standards*. Elsevier, Oxford, UK

Olfat H, Rajabifard A, Kalantari M (2010). Automatic Spatial Metadata Update: A new approach. *Proceedings of the FIG Congress*, Sydney, Australia

Poore B, Wolf E (2010). The Metadata Crisis – Can geographic information be made more usable? U.S. Geological Survey *Proceedings of the Second Workshop on Usability of Geographic Information*, London, 23rd March 2010. Available from: http://www.virart.nottingham.ac.uk/GI%20Usability/Workshop%20papers%20pdfs/ Accessed 1st March 2011

West L, Hess T (2002) Metadata as a knowledge management tool: supporting intelligent

agent and end user access to spatial data. *Decision Support Systems* 32, pp 247-264

Biography

Claire Ellul is a Lecturer in Geographic Information in the department of Civil, Environmental and Geomatic Engineering at University College London. Her research interests include spatial databases, data quality and metadata and approaches for handling large quantities of spatial data, in particular in 3D GIS.