Using Free and Open Source GIS to Automatically Create Standards-Based Spatial Metadata in Academia - First Investigations

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

[1]Dept. of Civil, Environmental and Geomatic Engineering, University College London, UK
[2] Laboratory of GeoInformation Science and Remote Sensing, University of Wageningen, The Netherlands

Abstract

The importance of understanding the quality of data used in any GIS operation has increased significantly as a result of the advent of Free and Open Source (FOSS) tools and Open Data, which in turn have encouraged non-specialists to make use of GIS. Metadata (data about data) traditionally provides a description of this quality information and permits data curation, but it is frequently deemed as complex to create and maintain. Additionally, it is generally stored separately from the data, leading to issues where updates to the data are not reflected in the metadata and to users not being aware that metadata exists. This paper describes an approach to address these issues in an academic context - tightly coupling data and metadata and automating elements of standards-based metadata creation and automating keyword generation and language detection. We describe research into the potential of the FOSS packages Quantum GIS and PostGIS to support this form of metadata generation and maintenance.

Keywords

QGIS, PostGIS, Metadata, INSPIRE, FOSS, Automation, Keyword generation, Language detection

1 Introduction

Advances in positioning, web mapping, mobile communications, Web 2.0 and Volunteered Geographic Information (VGI) (Goodchild 2007), along with the emergence of the Open Data movement, have led to increasing availability of spatial data (Budhathoki et al. 2008), with much of this data available free of charge (Coleman et al. 2009). The availability of free Geographical Information Systems (GIS) software (e.g. Google Earth, ArcGIS Explorer, Quantum GIS) encourages non-specialist users to make use of GIS tools and data.

In academia, this increase in available data and software, along with the requirement to curate the data, is coupled with a reduction in GIS expertise of the end user of such tools. Given this, having information to allow end-users to understand, manage and integrate the heterogeneous data they are using, and identify any limitations, becomes more important (Deng & Di 2009, Haklay & Weber 2008).

Traditionally, among GIS professionals, metadata ('data describing the data') has been created to curate data (Sboui et al. 2009). It details how data was derived, why it was captured, at what scale and how it has been processed, covering issues related to topological correctness, semantic, temporal and positional accuracy (Goodchild 2002, Longley et al. 2011, Van Oort 2005, Burrough 1994). It provides a formal description of the data quality (Kim 1999), allows for data reuse (Craglia et al. 2008) and avoids data duplication. Good metadata increases trust (Craglia et al. 2008) and helps increase the credibility of a dataset (Coleman et al. 2009). In general, therefore, "the purpose of metadata is to facilitate the interpretation of data" Sboui et al. (2009).

However, metadata is complex to create (Poore & Wolf 2010, Manso-Callejo et al. 2009, Batcheller 2008, Craglia et al. 2008) and is usually created by a dedicated team of professionals (Mathes 2004 in (Kalantari et al. 2010)). "Many view its generation as monotonous and time-consuming" (Batcheller 2008, p. 388). Standards are producer-centric (Goodchild 2007, Devillers et al. 2005) and quality may be variable (Rajabifard et al. 2009). Metadata production is often left to the end of a project, which results in metadata that is barely useful and often contains errors (West & Hess 2002). The current approach to data curation - where metadata is decoupled from the data it describes - further complicates this situation. Decoupled metadata may not be updated when data changes, and its existence is easily ignored by users.

This paper presents preliminary work on an approach to overcome these issues in the context of academic research and data curation. Using Free and Open Source (FOSS) GIS products - Quantum GIS 1.8.0 and PostGreSQL 9.2 with PostGIS 2.0 (to maximize potential uptake amongst academics without incurring licensing costs), we describe how metadata creation to the metadata standard used by INSPIRE (INSPIRE 2011*b*) can be, in large part, automated - in particular keyword generation and language detection. Importantly, this is done in a manner that tightly couples metadata and data.

The remainder of this paper is structured as follows: firstly, we briefly outline the importance of spatial data infrastructures and metadata in an academic context, considering the relevance of INSPIRE and the ISO 19115 standard, approaches offered by current vendors and previous attempts at automation. This is followed by an investigation into the automation potential of individual elements of ISO 19115 and a description of the system architecture used and implementation approaches taken. Results are presented, particularly for language and keyword automation and the paper concludes with a discussion and an overview of further work to be carried out.

2 Background

2.1 The INSPIRE Project

The INSPIRE (INfrastructure for SPatial InfoRmation in Europe) directive, issued by the European Union in 2007 (INSPIRE 2011*a*), sets up a framework for the creation of an European Spatial Data Infrastructure (ESDI), which will enable the sharing and comparison of environmental information among public sector organizations and facilitate public access to spatial information across Europe (INSPIRE 2011*a*). Data themes covered by INSPIRE are wide-ranging and include coordinate reference systems, addressing, administrative units, land cover, elevation, environmental monitoring facilities and natural risk zones.

As with any Spatial Data Infrastructure, metadata forms a core component of INSPIRE. For INSPIRE this is based on the ISO 19115 standard ISO (2003)(referred to as “INSPIRE metadata” in this document). Core elements of INSPIRE metadata cover resource identification, keywords, geographical location, temporal references, quality and validity of the data and information about the metadata itself (INSPIRE 2011*b*) along with issues relating to sourcing the data and licensing its use, as well as logical consistency (the degree to which the contents of the dataset follow the specification rules), completeness (are there gaps or missing data), positional accuracy, and lineage (how the dataset was acquired or compiled) (Goodchild 2007). Indeed, a total of 38 separate items of information can be identified INSPIRE (2011*b*).

2.2 Academic Context

The increasing availability of software and data is particularly relevant for many of the multi-disciplinary academic projects in which the authors of this paper are involved and which provide a motivation for the research described here. The power of GIS as a tool for the integration of data from diverse sources and disciplines means that it is frequently used in such projects. These projects, in turn, generate additional data. Curation of this data is an increasingly important area for academics, and is now mandated by funding bodies including the European Union FP7 FP7 (2011), the UK Environmental and Physical Sciences Research Council EPSRC (2011) and Economic and Social Research Council ESRC (2010). However, many academics do not have the skills required for such curation - and indeed may come from non-GIS disciplines as diverse as tourism studies, coastal geomorphology, anthropology, architecture and urban studies (Ellul et al. 2012).

Although it is as yet unclear whether the INSPIRE directive is specifically applicable to academia, and if so to what extent (Reid 2011), the general requirement to curate research data will most likely result in a requirement for the creation of standards-based metadata for academic datasets, to ensure interoperability and facilitate data exchange.

2.3 Metadata and GIS Software

Given that metadata has long formed an important element in the process of managing spatial data, and it is perhaps not surprising that many GIS packages provide functionality to create and maintain metadata as part of their functionality. The options offered by key packages are summarized in Table 1, along with metadata management tools provided by the INSPIRE project.

Package	**Summary**
ESRI ArcGIS 10.1	Metadata in ArcGIS is created via the separate catalog tool which lists available layers and datasets. Right clicking on a dataset opens up a properties window which allows the user to enter information including a title, tags, a summary, a description of the dataset, credits and access and use limitations. Metadata can also be imported from, or exported to, standards-based XML. It is embedded within the shapefile format, which means that it persists between projects. As the metadata is embedded, it is not possible to search through multiple metadata records unless these are exported.
Geomedia Professional 6.1	Geomedia Professional offers a 'catalog' tool for metadata creation and management, along with the capability to import existing metadata and export metadata for use in other systems. Catalog records are ISO 19115 compliant. Catalogs are stored as Microsoft Access databases (.mdb), decoupled from the datasets. The catalog creation process will automatically populate information including the bounding box in the Catalog Editor tool. No functionality is provided to search all created metadata for specific keywords or themes.

Package	**Summary**
Quantum GIS 1.8.0	In QGIS two alternative metadata options are offered. Users can create simple metadata directly with the properties of each dataset. The metadata is stored in the system project file (and therefore not available to other projects or users making use of the same datasets). Alternatively, a plug-in is also available - a metadata editor called Metatools, which can read and write metadata in ISO19115 format. The main purpose of this tool is to create standards-compliant metadata within QGIS, primarily for export to HTML format. This metadata editor is separate from the main workflow of a QGIS user (Lab n.d.). No functionality is provided to search all metadata.
INSPIRE Portal	The INSPIRE Geoportal (*The INSPIRE GeoPortal* n.d.) provides a central viewer for any available metadata created as part of the INSPIRE project. Unlike the stand-alone GIS packages, it provides a search tool for all created metadata, where searches can be spatial or text-based. Additionally, a metadata editor tool is provided to allow users to create INSPIRE compliant metadata. This can be validated using the tools provided. The metadata is held separately from the datasets referenced.

Table 1: Summarizing Existing Approaches to Metadata Handling

2.3.1 Limitations of Current Approaches

As indicated in Section 1, there are a number of issues with the current approaches. Firstly, the complexity of standards-based metadata means that users are not inclined to create or maintain it and therefore curate their data. Given the intricacy of metadata standards, even for specialists the complexity of creating and maintaining metadata is considered significant (Poore & Wolf 2010, Manso-Callejo et al. 2009, Batcheller 2008, Craglia et al. 2008). Secondly, metadata is, in most cases, de-coupled from the related dataset and in all cases does not form an integral part of the user's workflow when opening or editing data inside the GIS. This has two consequences - it is possible for users to use a dataset without being aware of any limitations or constraints - issues that are particularly relevant for novice users. It is also possible to edit and change the data without updating the corresponding metadata or to maintain metadata in one GIS but not make it available automatically to users of another GIS. This is particularly important to support interoperability.

To address these issues, metadata should be more closely coupled with the data itself and its creation should be as automated as possible. Where this is not possible metadata creation, maintenance and use should be integrated into the user's workflow. The remainder of this paper describes a first investigation into the potential of FOSS GIS to achieve these above aims.

3 Automating Metadata Creation

3.1 Previous Work

As has been seen (Section 2.3), a number of metadata elements are already created automatically by the various GIS packages. These include the identification of the bounding box coordinates of a dataset and the relevant projection or reference system. Beyond these basics, (Kalantari et al. 2010) have introduced a framework for the spatial metadata enrichment. Their work examines the potential of using concepts relating to tagging and folksonomies (collaborative tagging). Based on searches against a metadata repository, they assign the user's search words as keywords to any datasets that the user downloads as a result of a search process, and also propose direct user-tagging approaches to enriching metadata. (Olfat et al. 2012) introduce process-based metadata entry, which creates metadata in parallel with the dataset life-cycle, rather than after the generation of dataset or at the end of the project. They propose the coupling of metadata and data in one database. Their architecture is web based, making makes use of GML as a data transfer standard to support interoperability. Loose coupling is achieved by means of a layer of middleware (Olfat et al. 2012).

3.2 Automating INSPIRE Metadata Creation

A review of the metadata standard used by INSPIRE reveals that the population of a significant number of elements can be automated, in particular when the standard is applied in an academic context (which means that pre-defined, project specific, values can be used for some metadata). Indeed, it may be possible to automate the population of the majority of the mandatory elements of the standard. Table 2 outlines suggested approaches mandatory metadata elements (a full list, showing all metadata elements, can be found at `http://www.mapmalta.com/FOSS4G2013_FullTables.pdf`). The Table also describes which elements have been implemented in the prototype system described below (Section 4). At this point, it is important to note that, to work towards interoperability by allowing both data and metadata to be read by multiple GIS, it is assumed that all spatial data will be stored in a spatial database along with the corresponding metadata.

Metadata Element	Automation Potential	Mandatory	Implemented
Resource Title	Inserted manually. If not inserted by the user, default value is the dataset name.	Yes	Yes
Resource Abstract	Inserted manually	Yes	Yes
Identifier code and namespace (2 elements)	Take the Object Identifier of the spatial table in the database. This will form part of a unique URI for the dataset, which will also incorporate project and end user domain detail. In an academic context, a default value for a project or university could be used.	Yes	Yes
Resource type	Default to 'dataset'.	Yes	Yes
Resource language	It may be possible to implement this using language detection algorithms provided the dataset contains sufficient text. See Section 4.3.1 for details	No	Yes
Keyword(s)	This could be implemented by concatenating all text fields of the dataset and picking the top 10 repeating words while eliminating common words. See Section 4.2.2	Yes	Yes
Bounding Box (4 elements)	Can be automatically identified from the spatial coordinates in the dataset	Yes	Yes
Date of publication	Can default to the date that data was uploaded to the system, with updates when the data is edited. Manual verification required by the end user.	Yes	Yes
Date of last revision	Default to the date the data was uploaded to the system. Update automatically any time data edited	No	Yes
Date of creation	Default to the date the data was uploaded. Manual verification required by the end user	No	Yes
Limitations on public access and conditions of use (2 elements)	Given the academic context, a default value can be assigned, perhaps taking the most open value or perhaps on a per project basis.	Yes	No
Responsible party name, email and role (3 elements)	Based on user groups (identified from the user's login details and a corresponding lookup table).	Yes	Yes
Metadata contact name, email and date (3 elements)	This can be derived from the database login of the person uploading the dataset or creating the new dataset.	Yes	Yes
Metadata language	This can be detected by applying a language detection algorithm to the metadata (see 4.3.1 below)	Yes	Yes

Table 2: Potential Automation of some INSPIRE Metadata Elements

4 Implementing Metadata Creation in FOSS GIS

4.1 System Architecture

The approach described here builds on the concept of closely coupling metadata and data presented in (Ellul et al. 2012, Olfat et al. 2012). However, unlike Olfat et al. (2012), the coupling in this case takes place via triggers embedded in the database itself rather than relying on middleware. Triggers are automatic functions that run whenever data is inserted, updated or deleted in a database table, providing a very tightly coupled relationship between data and metadata, and their presence means that metadata is automatically updated when the dataset is edited in any way. Figure 1 shows the overall system architecture.

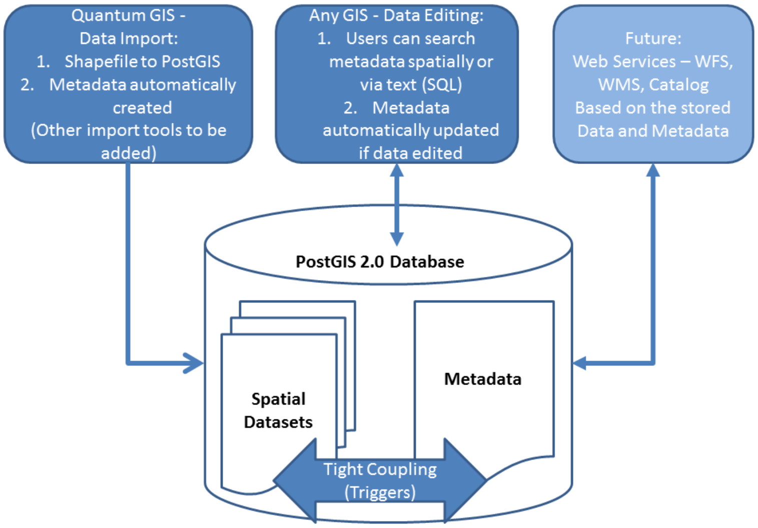
Quantum GIS -
Data Import:
1. Shapefile to PostGIS
2. Metadata automatically created
(Other import tools to be added)
Any GIS - Data Editing:
1. Users can search metadata spatially or via text (SQL)
2. Metadata automatically updated if data edited
Future:
Web Services – WFS, WMS, Catalog
Based on the stored Data and Metadata
PostGIS 2.0 Database
Spatial Datasets
Metadata
Tight Coupling (Triggers)

Figure 1: System Architecture

The system is built using Quantum GIS 1.8.0 and PostgreSQL 9.2 with PostGIS 2.0. The database was selected due to its interoperability with multiple GIS packages - including ArcGIS and Geomedia Professional. Quantum GIS offers the option to develop plug-ins using Python, and also offers excellent interoperability with the selected database.

4.2 Configuring the PostGIS Database and Triggers

To support metadata automation, two tables have been created in the PostGIS database - firstly, the metadata table itself and secondly, a table created to store required user information (based on the user's database login details) to automatically populate metadata when data is uploaded or modified. This 'database groups' table lists the group name (which corresponds to the PostGIS login group), the contact e-mail for that group and the user's organization name. s

4.2.1 Database Triggers

Triggers are created in PostgreSQL's inbuilt programming language - PL/pgSQL (*The PL/pgSQL Programming Language* n.d.). They add the following metadata details (see Section 4.3 below):

- The dataset extents, taken from the spatial extents of the data, using an ST_Envelope query
- The resource title, taken from the table name
- The Identifier Code, taken from the Object ID in the database
- The last revision date - defaults to the current date and time
- The metadata contact details - extracted from the PostGIS user's login details
- A bounding box geometry for the metadata, taken from the dataset extents
- The metadata date - taken from the current date and time
- The responsible party - taken from the PostGIS login
- Keywords - extracted from the dataset text (see Section 4.2.2)

A final procedure is then run to create a trigger on the newly uploaded dataset. This will automatically update the metadata every time the dataset is changed.

4.2.2 Identifying Keywords

The PL/pgSQL code used to identify keywords first identifies any text fields in the data table, and then splits the text into single words using a space as a delimiter. A UNION operation is used to generate a long, one-column, list of the words, and a *group by* query then used to identify the 10 most common words and known common words (*yes, no, or, and*) are eliminated from the list.

4.3 Writing the QGIS Plug-Ins

When using the Data Loading and Metadata Creation QGIS plug-in the user first selects a data file (shapefile) to load into the database, and an automatic connection to the database is established. The user can rename the file, and must type in the required metadata into the form - i.e. the title, abstract and, if available, lineage information. Once the user presses the 'OK' button, a Python process is run to load the data into the database and then insert the appropriate metadata.

The Python code first creates the spatial table in the database, naming the table with the filename of the uploaded file. The field names are identified from the shapefile and a geometry column of the appropriate type (again, identified from the shapefile) is added to the table. The SRID (spatial reference ID) value is taken from the form. Following this, the Python code iterates over the shapefile and inserts the data into the new table, row by row. Finally, the metadata language and dataset languages are identified (see Section 4.3.1 below), and required metadata inserted into the database. This metadata creation process - specifically the INSERT activity into the metadata table - triggers the PL/pgSQL procedures described above (Section 4.2.1) - i.e. the automated metadata creation.

4.3.1 Identifying the Metadata and Dataset Language

Both metadata and dataset language identification processes make use of a Python library known as “langid” (Lui & Baldwin 2012, n.d.). This language detection tool, based on machine learning algorithms (Lui & Baldwin 2011), is configured for 97 different languages, and works by a process of pre-training, in which common tokens for each language are identified through the examination of a set of documents in the given language. Importantly, it has been developed to allow cross domain applicability - for example, if the process has been trained to recognize Italian using a series of documents relating to air quality, it should be able to adapt this training to other Italian documents in a different domain. To detect the metadata language, three pieces of text are concatenated - the title, abstract and, if available, the lineage. For the dataset language, the first 10,000 characters of text are concatenated as the program is iterating through the dataset.

5 Testing Metadata Automation

In order to test the various key elements of the metadata automation process - and in particular keyword generation and dataset language detection, a range of Open Street Map datasets (OSM) (Haklay & Weber 2008) from ten different European countries were selected and uploaded into PostGIS using the plug-in described above (Section 4.3). OSM data was selected as it provides identically structured, multi-lingual (at least in part) data from around the world. This permits extended testing for keyword creation and language detection. An element of data cleaning has been carried out by Cloudmade. For each country the roads datasets (’highways’), location datasets (which detail key locations in each country), the points of interest datasets and the administrative boundary datasets were downloaded from `http:\\www.cloudmade.com`. Table 3 shows an extract of the results obtained, with the number of keyword occurances given in the brackets in each case.

Dataset	**ResourceLanguage**	**Keywords**

Dataset	ResourceLanguage	Keywords
austria_administrative	de	8(10889), 6(2158), 9(893), 10(875), 2(690), 4(527), /(371), Border(279), 7(264), StraÃ?e(203)
austria_highway	ht	track(288357), residential(265675), service(179314), path(86165), unclassified(78983), footway(78193), tertiary(32408), secondary(29048), StraÃ?e(28442), primary(24575)
austria_location	de	MÃijnchen(21504), hamlet(19230), Wien(13821), village(13086), 1(5866), Germering(5317), 82110(5312), 2(5292), Bad(4812), locality(4628)
austria_poi	en	Public(122871), Services(119379), Government(119379), Power(60686), Tower(59271), Automotive(52389), Tourism(45779), Bus(24174), Eating and Drinking(22032), Parking(20230)
greece_administrative	el	8(842), 6(317), 4(293), -(149), 2(95), Border(70), Î(60), sÄśnÄśrÄś(54), Î?ÎżÎżÎňÎť Îś(40), 7(38)
greece_highway	la	residential(107535), tertiary(22959), track(17127), unclassified(8819), secondary(8291), primary(6817), ?d??(6644), service(6377), footway(4517), motorway(3605)
greece_location	qu	village(6125), hamlet(3021), ???????(1344), ????????(1225), ??????(1130), ?????(1103), ?????????(884), ??????????(526), ????(521), town(298)
greece_poi	en	Public(20481), Services(20226), Government(20226), Tower(14798), Power(14757), Automotive(7226), Tourism(4412), Pedestrian(3117), Crossing(3117), Eating and Drinking(2752)

Table 3: Results for Open Street Map Datasets

5.1 Comparing the Results

Table 3 above gives a sample of the results obtained[1]. The results show that metadata has been correctly identified as being in the English language in all cases. However, in general both the keyword extraction process and the language identification process for the resource have yielded mixed results.

Firstly, all the points of interest data yielded 'English' as the language and terms including 'Public', 'Services', 'Tourism' and so forth as keywords. Examining the OSM points of interest datasets yields a potential explanation. Much of the data is in fact placed into category names which are given in English (no matter the country in question) - categories include 'Automotive', 'Government and Public Services', 'Tourism' and so forth. Additionally, although perhaps less expected, the types of the points are also given in English - for example in the Austrian dataset, we find 'Museum:Ortsmuseum Tutzing' and 'Significant tree' and 'Peak:Oberer Burgstall'. If the language issue is perhaps put to one side (as the

[1] Note that data is only shown for Austria and Greece due to space restrictions. A full listing can be found at `http://www.mapmalta.com/FOSS4G2013_FullTables.pdf`

data is, indeed, in English) it could be said that the keywords do provide a good representation of the Points of Interest Dataset, giving a mix of the types of information that this dataset provides.

For the administrative data, in all cases except for Belgium, the correct language was identified. In the Belgian case, however, the language was identified as 'lb' - Luxembourgish. This could be due to the fact that in Belgium both the French and Flemish languages are used. Keyword identification, on the other hand, was not as successful - indeed, numbers were identified as the most common 'words' in all cases. Examining the datasets again yields an explanation - in all the datasets, much of the 'name' data is blank (or null), whereas the 'administrative_level' data - which is a number detailing where on the overall Administrative boundary hierarchy the data element falls - is fully populated.

For the highway data, a more mixed result is noted - Malta, Italy, Spain and Greece all yielded 'Latin' as the language, Portugal and Sweden correctly yielded Portuguese and Swedish. However, the Netherlands yielded 'ms' (Malay), Belgium yielded 'jv' (Javanese) and Austria yielded 'ht' (Haitian). Examining the Belgian data, it can be seen that it is a mixture of Flemish, French and English, although English terms such as 'residential' and 'path' do dominate. Keywords in this case were given predominantly in English (due to the underlying data) and included 'track', 'unclassified', 'residential', 'footway' and 'cycleway'. Some common words in each language also made it to the list - Triq (Maltese), Via (Italian), Rua (Portuguese), Calle (Spanish), Strasse (Austria), Rue (Belgium), $O\delta o\varsigma$ (Greece) which all translate as street.

Finally, for the location dataset language identification, Sweden yielded 'Danish', both Spain and Portugal yielded 'Spanish', the Netherlands and Belgium yielded 'Netherlands', Malta yielded 'Latin', Italy yielded 'French' and Austria yielded 'German'. Greece yielded 'qu' (Quecha, which is a native South American language). In this case, the latter could perhaps be attributed to problems with the Greek characters in the text, which rendered as "?" in the database, along with the inclusion of Greek place names transliterated into English (such as 'Komianata' or 'Agii Deka'). Again, keywords were predominantly in English - 'locality','hamlet','village' due to the English language place data embedded in the datasets. However, these were mixed with place names (Aachen, Fgura, Birmingham, Munchen, Wein, Brugge, Trento) representing the most commonly used location points in each dataset. Provided the user understands English, the keywords do to a certain extent represent the dataset well.

To summarize the above results, in three cases out of the four tested, the keywords yielded from the datasets did provide a relatively useful list of words relevant to the dataset in question. The results for the fourth case - administrative boundaries - could perhaps be improved upon by eliminating keywords having very short length, or consisting of numbers, from the potential keywords list. Equally, it is important to identify and remove all common words ('and','or' and so forth) in the relevant languages from the list of potential keywords. This was temporarily hard-coded for the English language, but would require input from speakers of other languages to add to this list, which should then be stored in a table in the database.

Language identification also yielded rather mixed results, in particular where multiple languages were included in the dataset. A number of heuristics could be suggested to improve this process, however. A simple spatial intersection with a world map would identify the country where the data is located. This could then be used as a suggested language or languages to the language identification process. This is particularly important when two languages are close in nature - e.g. Flemish, French and Luxembourgish. The possibility of multiple languages within one dataset should also be considered and accounted for in the data model.

6 Discussion and Further Work

The work described here details a preliminary investigation into the potential of automating metadata - and in particular an investigation into the potential of closely coupling metadata and data, automatically generating keywords and detecting the language used in the dataset and metadata. Despite the issues with language identification, overall the work yielded promising results. Importantly, we have shown that metadata and data can be tightly coupled so that modifying data automatically updates the metadata. We have also shown that this is possible within FOSS GIS software. By embedding the coupling within the spatial database, the functionality to maintain dataset and metadata synchronized when data is edited is interoperable across multiple FOSS and non-FOSS GIS platforms. We have been able to automatically populate a total of 18 of the 20 mandatory INSPIRE fields, with a further 4 optional fields populated (from a total of 18 optional fields).

Both datasets and metadata are stored in an open spatial format, which means that they can be shared with other GIS packages - this makes the entire metadata catalog searchable within the GIS. The central database approach also permits the data and metadata to be published as Web Feature Services and Catalog Services for the Web, providing a discovery type service similar to that used in INSPIRE. While the work described in this paper has focused on metadata in an academic context many of the approaches described above are relevant elsewhere, although it is possible that using these approaches in a more general context may reduce the number of metadata elements that can be automated.

A number of technical issues remain to be addressed - in particular, the difficulty encountered when handling Greek and other non-Latin characters in the database and potential performance issues caused when significant changes are made to the underlying datasets on a row by row basis. Potentially, the user could be given the opportunity to temporarily disable metadata updates, or to set them to run in batch mode overnight.

Increasing the interoperability of the approach also presents an interesting challenge. At this point in time, functionality developed works when a pre-created dataset is imported into the database, and when the resulting dataset is edited by any GIS the automated elements of the metadata are updated - i.e. partial interoperability has been achieved. Interoperability should, however, be extended to incorporate any new spatial table created in the database or imported via other mechanisms. Semantic interoperability of the manually-populated elements of the metadata standard also present a problem and may require the development of plugins for the creation of the manual elements of metadata via other GIS or import of existing metadata. It would also mean that keyword and language identification would not necessarily be immediately possible, but could instead be run as a batch process once sufficient data is added into the table. For total interoperability, the language detection algorithm should be incorporated directly into the database rather than embedded in the plug-in.

Given the mixture of languages within the OSM datasets, perhaps the next step in this research should also be to identify appropriate single-language datasets (along with corresponding metadata) to conduct further tests. Language experts for each language should also be involved to ensure that the results yielded are appropriate - in particular the keywords identified. Once this is complete, further work could continue on the other elements of metadata automation. The system could also be extended to include non-INSPIRE metadata - (Ellul et al. 2012) notes a number of elements such as tags, dataset and metadata ratings that could be relevant. Potentially, given that the aim of this tool is for use in an academic setting, additional project-related information (for example which Work Package generated a dataset) could also be added. Finally, issues relating to deployment should also be considered - these tools require a level of expertise to initially set up and configure, which could be provided by data management support staff within each academic institution.

References

Batcheller, J. K. (2008), 'Automating geospatial metadata generationŮan integrated data management and documentation approach', *Computers & Geosciences* **34**(4), 387–398.

Budhathoki, N., Bruce, B. & Nedovic-Budic, Z. (2008), 'Reconceptualizing the role of the user of spatial data infrastructures', *GeoJournal: An International Journal on Geography* **72**, 149–160.

Burrough, P. (1994), *Principles of Geographical Information Systems for Land Resources Assessment*, Oxford Science Publications, Clarendon Press, Oxford, UK.

Coleman, D., Georgiadou, Y. & Labonte, J. (2009), 'Volunteered geographic information: The nature and motivation of produsers', *International Journal of Spatial Data Infrastructures Research* **4**, 332–358.

Craglia, M., Goodchild, M., Annoni, A., Camara, G., Gould, M., Kuhn, W., Masser, I., Maguire, J., S, L. & 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.

Deng, M. & Di, L. (2009), 'Building an online learning and research environment to enhance use of geospatial data', *International Journal of Spatial Data Infrastructures Research* **4**.

Devillers, R., Bedard, Y. & Jeansoulin, R. (2005), 'Multidimensional management of geospatial data quality information for its dynamic use within GIS', *Photogrammetric Engineering and Remote Sensing* **71**, 205–215.

Ellul, C., Winer, D., Mooney, J. & Foord, J. (2012), 'Bridging the Gap Between Traditional Metadata and the Requirements of an Academic SDI for Interdisciplinary Research', *Spatially Enabling Government, Industry and Citizens* .

EPSRC (2011), 'EPSRC Policy Framework on Research Data', `http://www.epsrc.ac.uk/about/standards/researchdata/Pages/default.aspx`.

ESRC (2010), 'ESRC Research Data Policy', `http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf`.

FP7 (2011), 'Research Data Management - European Commission - FP7', `http://www.admin.ox.ac.uk/rdm/managedata/funderpolicy/ec/`.

Goodchild, M. (2002), 'Introduction to part i: Theoretical models for uncertain GIS', *Spatial Data Quality* pp. 1–4.

Goodchild, M. (2007), 'Citizens as sensors: the world of volunteered geography', *GeoJournal* **69**, 211–21.

Haklay, M. & Weber, P. (2008), 'Openstreetmap - user generated street map', *IEEE Pervasive Computing* pp. 12–18.

INSPIRE (2011*a*), 'About INSPIRE - Infrastructure for Spatial Information in Europe', `http://inspire.jrc.ec.europa.eu/index.cfm/pageid/48`.

INSPIRE (2011*b*), 'INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119', `http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf`.

ISO (2003), 'ISO 19115:2003, Geographic Information – Metadata'.

Kalantari, M., Olfat, H. & Rajabifard, A. (2010), 'Automatic spatial metadata enrichment: reducing metadata creation burden through spatial folksonomies', *Global Spatial Data Infrastructures 12 World Conference: Realising Spatially Enabled Societies* .

Kim, T. (1999), 'Metadata for geo-spatial data sharing: a comparative analysis', *The Annals of Regional Science* **33**, 171–181.

Lab, G. (n.d.), 'Working with metadata using metatools for QGIS', `http://gis-lab.info/qa/metatools-eng.html`.

Longley, P., Goodchild, M., Maguire, D. & Rhind, D. (2011), *Geographical Information Systems and Science Third Edition*, Wiley Hoboken NJ.

Lui, M. & Baldwin, T. (2011), 'Cross-domain feature selection for language identification', `http://www.aclweb.org/anthology-new/I/I11/I11-1062.pdf`.

Lui, M. & Baldwin, T. (2012), 'langid.py: An off-the-shelf language identification tool', `http://www.aclweb.org/anthology-new/P/P12/P12-3005.pdf`.

Lui, M. & Baldwin, T. (n.d.), 'langid.py', `https://github.com/saffsd/langid.py`.

Manso-Callejo, M., Wachowicz, M. & Bernabé-Poveda, A. (2009), 'Automatic metadata creation for supporting interoperability levels of spatial data infrastructures', *GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges* .

Olfat, H., Kalantari, M., Rajabifard, A., Senot, H. & I, W. (2012), 'Spatial metadata automation: A key to spatially enabling platform', *International Journal of Spatial Data Infrastructures Research* **7**, 173–195.

Poore, B. & Wolf, E. (2010), 'The metadata crisis - can geographic information be made more usable?'.

Rajabifard, A., Kalantari, M. & Binns, A. (2009), 'SDI and Metadata Entry and Updating Tools', *SDI convergence: Research, Emerging Trends and Critical Assessment* .

Reid, J. (2011), 'The EU INSPIRE Directive and what it might mean for UK Academia', `http://www.data-archive.ac.uk/media/338363/INSPIRE7Oct2011jr.pdf`.

Sboui, T., Salehi, M. & Bédard, Y. (2009), 'Towards a quantitative evaluation of geospatial metadata quality in the context of semantic interoperability', *6th International Symposium on Spatial Data Quality* .

The INSPIRE GeoPortal (n.d.), `http://inspire-geoportal.ec.europa.eu`. Accessed: February 2013.

The PL/pgSQL Programming Language (n.d.), `http://www.postgresql.org/docs/9.0/static/plpgsql.html`.

Van Oort, P. (2005), 'Spatial data quality: From description to application', *PhD, Wageningen University, The Netherlands* .

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**, 247–264.