Multipurpose Metadata Management in gvSIG

Laura Díaz, Michael Gould, Arturo Beltrán, Alejandro Llaves, Carlos Granell

Department of Information Systems, Universitat Jaume I, Castellón, Spain
{laura.diaz, gould, carlos.granell}@uji.es

Abstract

*Metadata exist in multiple forms and for multiple purposes, however most of the attention in the geographic information community is focused on publication of external metadata in catalogues. This is important for facilitating geospatial data discovery and for data sharing in the SDI context, however system-specific metadata for internal data management is also important for optimizing the use and value of a GIS desktop application. The gvSIG project (*www.gvsig.gva.es*) has decided to expand the definition and use of metadata, to not only integrate metadata creation and export within the desktop client (as opposed to using a stand-alone metadata editor) but also to embed metadata into the internal data structure in order to improve the management of both data and processes on the data.*

We present an on-going development of a multipurpose metadata manager for documenting well-known imagery and cartographic data sources, being implemented within an open-source software GIS/SDI client. The first prototype includes a metadata manager capable of semi-automatic extraction of explicit metadata from data resources for both internal metadata for user efficiency purposes and external metadata to be catalogued for data discovery in an SDI. The graphical editor displays metadata to the user who can edit it and export it to well-known formats. The nature of the integrated workflow facilitates metadata creation and management, hopefully contributing to a change in mindset as to the cost/benefit ratio of generating and exploiting metadata, a necessary ingredient for successful Spatial Data Infrastructures and also for internal data use and management.

Keywords: metadata, metadata manager, metadata extraction, service catalogues, gvSIG.

1. Introduction

Initiatives such as the EU INSPIRE Directive (INSPIRE, 2007) are making regional and national Spatial Data Infrastructures (SDI) not only desirable but legally required. INSPIRE mandates the creation and maintenance of metadata and related discovery services, which often are the first visible value-added element of a SDI. Metadata are necessary to allow description of data and service resources, becoming a key element for judging data fusion possibilities and for user discovery –do the data I need exist and, if so, where and under what conditions?-- in an SDI (GSDI, 2008; Granell et al, 2008). There are three basic roles for metadata: resource discovery, resource evaluation and facilitating resource consumption by human or machine. Metadata has been defined traditionally as “data about data” however (Nogueras et al, 2005) extends this to define them as

"data plus context for its use (documentation)". In the project reported here we have been experimenting with this context, in its various guises.

It is normally the user who must manually create these metadata. This process is currently undertaken using simple text editors and outside of the GIS or image processing workflow. This documentation process can and should be automated to the extent possible, given that informatics technology has greatly improved since the early days (1980s) of the digital libraries that gave birth to the current manual metadata creation methodology. A few proprietary metadata extraction solutions have appeared, however in most cases their workflow is restricted to creation and cataloguing using client and server software from the same commercial family, whereas SDI related initiatives such as INSPIRE and GMES are promoting heterogeneity and interoperability, making the availability of open source solutions and non-proprietary formats all the more attractive.

These traditional geo-metadata, for example those following the ISO 19115-39 standards, are what we term external metadata, created for others to use. Several researchers have cautioned against trying to use these same external (general discovery) metadata also for internal purposes, as the overloading can cause confusion and/or unnecessary complexity. Therefore the gvSIG solution also creates and stores internal metadata for user queries, history tracking, etc., for efficiency purposes. Only when data sharing is desired does the user initiate the metadata export process, passing the necessary metadata through stylesheets to create SDI-compatible external metadata records, according to different supported metadata standard formats (ISO, FGDC, Dublin Core). Furthermore, these records can be exported semi-automatically to GeoNetwork catalogue services (other catalogues will be supported in future phases of the project) without leaving the GIS desktop working environment.

In this sense we have produced a GIS which is also a full-function SDI client, facilitating discovery and sharing of geospatial data in addition to local geoprocessing. GIS/SDI users are able to document their new data resources at the time of creation (rather than later on, when details may have been forgotten), or at least while viewing and utilizing the data, directly within the normal workflow without the need to work with the geodata in one application and text descriptions in another. The GIS/SDI client user can view the metadata describing selected resources and make use of it internally to increase efficiency and avoid for instance task duplication, can edit that metadata, and finally can publish the metadata for external use in the wider community.

2. Project Context

A large project for migration from proprietary to free software, initiated in 2004 by the Valencia regional government (Generalitat Valenciana), has produced a client software product called gvSIG[1]. What began as a simple, Java-based GIS client quickly evolved to become a full-function, component-based SDI client, implying that it facilitates discovery and sharing of geospatial data in addition to local geoprocessing. With this SDI-based data sharing in mind we have designed a gvSIG extension (component) to semi-automatically extract metadata from well-known geodata formats, system information and user preferences, to manage metadata for multiple purposes. On

[1] http://www.gvsig.gva.es

one hand we create standard metadata for discovery purposes in SDI environments and on the other hand we create specific metadata for localized application efficiency. The idea, as stated previously, is that GIS/SDI users are given the capacity to visualize and manage the extracted metadata of their new data resources at the time of creation, or at least while viewing and utilizing the data, directly within the normal workflow without the requirement to work with the data in one application, and text descriptions in another, which is both cumbersome and leads to synchronization problems when updates are made to either data resource or description.

In this line we manage two concepts when talking about metadata. First we talk about internal metadata, metadata that are read from the data source, system, user preferences, etc. and it is use for internal use within the GIS application. Examples might be the number of features in the dataset, when each was last modified, and by whom. The second concept is external metadata, which includes a minimum metadata set (for example the ISO 19115 core) sufficient to allow resource discovery (data sharing) in SDI environments. We regard these two metadata concepts, and their use cases, to be fundamentally different. As related work we can find several open source implementations for external metadata creation such as the approach taken by CatMDEdit[2] which is a metadata editor tool that facilitates the documentation of resources and additionally can visualize some data file formats. The Cataluña SDI project produced the MetaD stand-alone editor, which is freeware however not open software; similar solutions are provided by members of the SDI community in the United States (FGDC provides a list of metadata editors. Other solutions are metadata editors integrated in Catalogue Services such as terraCatalogue[3] and GeoNetwork[4]. Almost all these approaches are separated metadata editors where metadata are created manually to be later stored in Catalogue Services. Our approach differs in several points. First, we provide editing capabilities within a stand-alone GIS package, where the technician can edit the metadata while he is working with the data; these editing capabilities are supported by a manager component which in the background is collecting explicit metadata contained in the data itself and other system and user information to minimize repetitious manual creation of the entire metadata record. Second, the metadata management within gvSIG, as SDI client, will allow users to directly export metadata to Catalogue Services once the metadata format is validated. Third, as mentioned before, gvSIG metadata manager will create and manage not only these kind of metadata, but also internal metadata, that although these metadata are not of general interest to be published in catalogues they are very useful to guarantee efficient workflow while managing spatial data in the client. In the future we plan to package this internal metadata in the same package exported to catalogues; the receiving user can decide to try to use it or ignore it.

Recently, Google and several multimedia information retrieval projects have demonstrated that data resources may be encountered without the need for tedious manual data product documentation, thanks to intelligent methods for intuitive metadata extraction from the data source.

[2] http://catmdedit.sourceforge.net
[3] http://www.sdi-suite.de/en/terra-catalog.shtm
[4] http://sourceforge.net/projects/geonetwork

This is the direction we have chosen to follow (Gould et al, 2006) in future phases of the gvSIG metadata extension project.

Beard (Beard, 1996), in a workshop about environmental geospatial data, summarizes the methodologies to collect metadata:

1. manually generated
2. *look-up* in a reference table
3. value measurement
4. metadata computation
5. metadata inference

The first methodology has been described earlier: the user edits the metadata (xml format) by using a metadata editor like MetaD or CatMDedit. This is the default methodology of SDIs today.

The second methodology supposes that a metadata element is created by matching with other metadata, for example obtaining a toponym through a gazetteer service using the four coordinates of a bounding box.

The third methodology supposes that during the collection of geodata certain values can be observed and added to a metadata automatically, such as in the case of reading data from a weather web service connected to a thermometer sensor device. Furthermore the season of the year might be inferred by the temperature.

The fourth methodology calculates metadata from geodata, such as determining the province of a village automatically using administrative hierarchies in a database.

The fifth methodology is to infer metadata from other metadata or data. According to Beard (already a decade ago) this is the best method- in fact in some situations is the only method- for metadata creation post hoc, i.e., documenting already existing metadata. This would be the case inferring metadata period by a temperature value collected by methodology three, for example a rule would establish that a temperature under 15ºC in Tenerife means season=winter. Beard also points out that metadata inference overlaps with the data mining and information retrieval fields, a fact also noted in (Goodchild, 2007). These inferred metadata are what we call implicit metadata.

3. Metadata Types and Purposes

Internal metadata, although not designed to be published in catalogues for SDI sharing purposes, is very useful to guarantee an efficient workflow while using GIS. These metadata will not be edited by the user however they may be visualized by her. Its main purpose is to help monitor the data status for the sake of efficiency and performance.

Traditional external metadata may overlap in certain elements with the internal metadata, in which cases the user need not create them using manual editors, because they have been automatically created by the metadata manager, the remaining elements needed to fulfil a given standard core may be added by the user using the metadata editor included in our solution integrated in the GIS package, and the editor will validate and export the metadata record to other formats and then upload to catalogue services.

We believe that methodologies and the architectural status quo defining current SDIs are not sufficient to allow for the massive scalability needed by SDI version 2.0 (following web 2.0 terminology). This is mostly due to the current need for users to edit text data, a task that will not scale when we wish to document not only a few datasets but hundreds of millions of features, services, and processes. Additionally SDI v2.0 will need to store and provide access not only to official resources but also to the contributions of the 6000 million popular sensor platforms (humans) moving around the world, carrying potentially sensors (Goodchild, 2007).

In this transition from v1.0 to v2.0 the objective is to maintain the basic behaviour of a current SDI, improving underlying implementation to augment efficiency and simplicity. One challenge for SDI v2.0 will be convert the collection and exploitation of metadata to become a transparent, scalable and less tedious part of the overall workflow. Current manual metadata editing workflows are tedious and under-utilized, resulting in a scarcity of useful metadata in Spatial Data Infrastructures, as pointed out in (Craglia et al, 2007). If the next generation SDI were to follow more closely recent advances in the allied fields of multimedia, it will be automatically extracted during the production and usage process of the data (see for example Bulterman, 2004; Gould, 2005) as is the case for digital cameras for example.

Our solution seeks to minimize user effort, automating metadata creation to the extent possible, and so in the following section we review possible methods to extract and collect metadata.

3.1 Extracted Metadata

Most of the geospatial open source libraries read and write multiple data formats, GDAL/OGR being one of the most popular because of the ample list of supported formats. Using these libraries it is easy to extract certain file-based information to create metadata. Additionally, the operating system provides us with further relevant information regarding dates, files, users, etc. Manso et al (2004) have reported on previous efforts to automatically extract interesting subsets of standard metadata records.

3.2 Inferred Metadata

We can infer metadata from other metadata or from geodata by using data mining techniques, information retrieval and use of metadata context, reasoning techniques, etc. Huge gains may be made here by following so-called Google methods, exploiting probabilities to determine, for example, that if the file's legend contains soil classes then the "map" is likely a soils map; if the file is served by the Cartographic Agency of Andalucia then the area of interest is probably in the region of Andalucia; if the features are points then the theme is likely NOT lakes or parcels. These may seem like obvious, almost silly examples, yet current metadata and catalogue solutions are not exploiting them. Catalogue services also do not normally track user data, and therefore, they do not "know" where the user is from, and that he/she almost always requests data for Spain, and therefore a search on Valencia is probably NOT referring to Valencia Venezuela or Valencia California.

3.3 Search By Crawler

The WWW crawlers use optimized searches by indexing. These indexes are formed by *crawlers* that continuously crawl resources connected to the net. Related to the inferences previously mentioned, recent advances in automating searches and on reference- or link-following need to be further explored and tested.

The key to a future success is likely to be in the combination of these methodologies, which first requires a refocus on the part of the SDI community, from a cartographic focus to an information systems focus. To cite an example (Forsyth and Wilinsky, 2003), experts in libraries but not versed in new technologies have recently pointed out: "This class of strategies is likely to be successful only for metadata that can be inferred from the object itself. For example, it may be possible to determine the names of those present from a picture, but it is likely to be impossible to determine the time and data at which the picture was taken". This is absurd from an IT perspective: today most the digital cameras have internal metadata created by the camera itself to indicate for instance date and time and even place, for a good example see: http://es.wikipedia.org/wiki/Imagen:Geotagging.png

4. Metadata Management Platform

The ongoing work is to become available in 4Q2008 as a pilot plug-in of gvSIG. The metadata manager architecture is shown in figure 1. It interacts with the gvSIG core to handle the metadata associated to all the gvSIG resources pointed up to be described with metadata, whether for internal use or for being exported to an standard format or published in a catalogue service.

The metadata manager will be working in the background annotating all the metadata while technicians are working with their geospatial data. gvSIG will be using the internal metadata to avoid task duplication or recalculations, stored intermediate results, etc. At any moment the metadata can be visualized by the user. In case the user wants to share metadata or publish metadata in standard formats, the metadata manager will warn about the status of the metadata, i.e., the user could edit the metadata to fulfil a minimum set of a standard format. Included in this workflow there is a *publisher* module, in a form of a user-friendly wizard to guide the user to publish these metadata in a catalogue service.

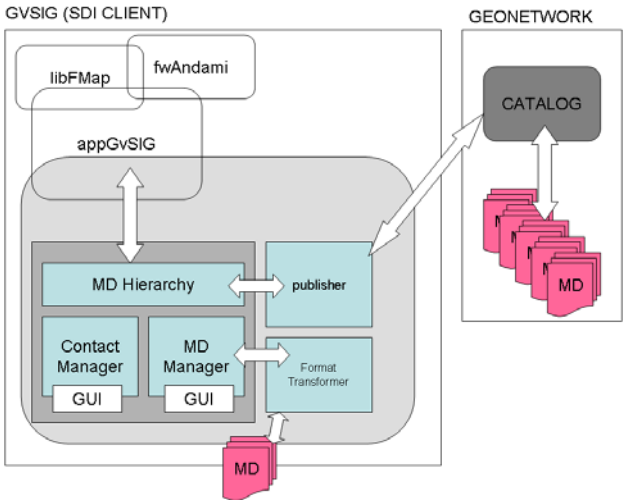

Figure 1. Metadata Manager Architecture

We began by defining within the central structure of gvSIG, an internal metadata object which would store various types of metadata: internal, external, and possibly user-defined. The various metadata elements collected are stored in an XML format file. Nearly all popular metadata formats such as ISO 19115 and Dublin Core include elements describing the author of the dataset. Therefore we adapted the gvSIG user configuration panel, to accommodate stable user-related metadata including personal, work-related and professional contact information to later be used, where relevant and when permission is granted, in the metadata collection process. Given the above preparation, let us look at a typical use case. A technician using gvSIG has combined basic geodata including terrain data such as slope and aspect, with vegetation data, to create a rough forest fire risk map. Assuming she has permission to share this new dataset, she then undergoes the process of publishing the risk map to a map server, and would also like to (or should be required to) also publish its description to a metadata catalogue service such as that currently available at the European Geoportal (http://www.inspire-geoportal.eu/).

At the moment a well-known format geodata file is opened in gvSIG, we check for its metadata object (internal data store) and if it does not exist we create one. Then we automatically extract so-called internal metadata (format, resolution, spatial reference system, creation date, etc.) using the operating system, GDAL/OGR (Geographic Data Abstraction Library)[5] and related packages, and we add these metadata to an internal metadata object. Here we have adapted metadata extraction methodology described by (Manso et al, 2004). The user has the option to open this metadata file and to add, using an integrated metadata editor, additional textual information (such as Abstract) that might be required by standard metadata formats as defined by organizations such as FGDC or ISO TC211. In the case of our use case, the resulting dataset, risk map, is assigned a metadata object and the process is as described above. The final step in the workflow is that when and if the user decides to publish the metadata record to a catalogue service (we have used FAO's GeoNetwork open source) the gvSIG metadata manager checks the validity of the metadata present in the MDML file, the validation will depend on the metadata standard that has been chosen to publish, thus the standards that the Catalogue Service supports. If the metadata conform to minimum requirements according to the output format/standard selected, then the metadata manager uses stylesheets to generate an XML file compatible with the catalogue service.

5. Current Functionality

5.1 Documentation Levels

Although not all the resources managed in gvSIG are relevant to be described with external metadata to be published in SDI, there is a multi-level hierarchy that permits us to create metadata for internal use for multiple resources like layers, features, processes, maps, etc. New plug-ins installed in gvSIG will be able to register new metadata objects to be handled within the metadata manager.

[5] http://www.gdal.org/

5.2 Metadata Extraction

Currently different techniques are developed to extract automatically metadata from data sources, gvSIG drivers using other open sources libraries are able to read file format headers and other information to collect metadata for different uses, besides these the metadata manager extracts and collects metadata from the system, user preferences and handles user contact information as well. Future phases (2009) will include inference and information retrieval techniques and to create metadata, so the user will hardly have to edit or add metadata to publish it in SDI, thus facilitating the proliferation of metadata and thus the resource discovery in distributed information platforms.

5.3 Internal Storage

The metadata manager stores the metadata internally to avoid task duplications, the metadata are stored in an internal database, ready to be exported to standard formats when required by the user.

5.4 Visualizing and Editing Metadata

Only what we termed external metadata are in fact visualized and edited by the user, in this case, when a user wishes to edit a metadata to export it or publish it in a catalogue service, he will choose a standard format for this purpose. Once it has been chosen, the metadata manager will start a wizard that will guide the user to view and edit the metadata according to the format and validating the metadata fields.

5.5 Importing and Exporting Metadata

The wizard mentioned in the point before will guide users to import and export metadata validating it according to a standard format to be shared by multiple users without having to publish it in a Catalogue Service.

5.6 Publishing Metadata in Catalogue Services

As mentioned before, the metadata that has been semi-automatically extracted and collected from different sources by the metadata manager can be published in Catalogue Services. For this purpose Metadata Manager includes the publish module to guide the user through the steps from the choice of a standard format and validation until the final publishing in an instance of Catalogue Service in an SDI.

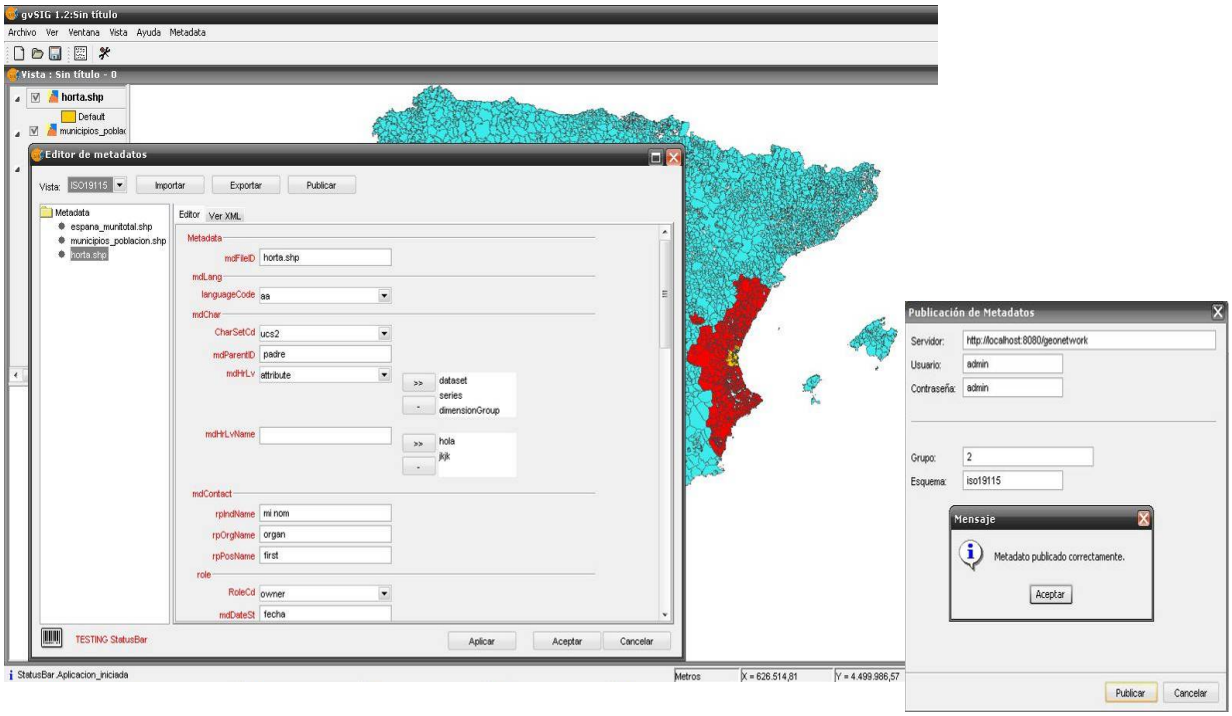

Figure 2. Screenshot of the Pilot Metadata Editor, With Spanish User Interface

6. Conclusions

This implementation of the concept of semiautomatic extraction and management of metadata facilitates the creation and edition of images and geospatial data to be published in a Spatial Data Infrastructure. The integrated nature of this solution within the user workflow hopefully will lead to a proliferation of metadata creation, thus improving the functionality and value of SDIs.

The Metadata manager within gvSIG handles both internal metadata and external metadata, which sometimes overlap, providing added value to the user: while internal metadata provide increased efficiency in the user workflow, this kind of metadata supports the creation of the external metadata that the user will require when sharing data in SDIs.

More interesting future developments will consist in more intelligent methods to extract metadata by using inferential reasoning techniques from other metadata and data associated. Intuitive extraction of intrinsic (context-based) metadata of the data source in Google-like techniques, including deductive methods to create well formed free text. Other research lines about external metadata are for instance the use of indexing techniques to find data using simple metadata instead of creating complex standard formats stored in Catalogues.

7. Acknowledgements

This work has been partially funded by the Generalitat Valenciana and FEDER funds in the framework of the gvSIG project.

References

Beard, K. 1996. “A Structure for Organizing Metadata Collection”, The Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, New Mexico. NCGIA, University of California at Santa Barbara.

Bulterman D. 2004. Is it Time for a Moratorium on Metadata? IEEE Multimedia, October-December, 10-17.

Craglia, M, Kanellopoulos, I and Smits, P. 2007. Metadata: where we are now, and where we should be going. Proceedings of 10th AGILE International Conference on Geographic Information Science 2007. Aalborg University, Denmark

Díaz, L, Martín, C, Gould, M, Granell, C and Manso, M.A. 2007. “Semi-automatic Metadata Extraction from Imagery and Cartographic data”. International Geoscience and Remote Sensing Symposium (IGARSS 2007). Barcelona, Julio 2007. IEEE CS Press, pp. 3051-3052.

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union, 25 April 2007. http://inspire.jrc.it/

Forsyth, DA, Wilensky, R. 2003.“Research issues for digital libraries”, NSF Post-DL Futures Workshop, Chatham, MA. http://www.sis.pitt.edu/~dlwkshop/paper_wilensky.pdf

Goodchild, M. 2007. “Citizens as Voluntary Sensors: Spatial Data Infrastructure in the

World of Web 2.0”, International Journal of Spatial Data Infrastructures Research. Vol. 2, 24-32.

Gould, M, Rocha, J, Nativi, S, Nogueras, J and Manso M.A. 2006. “Near-term metadata challenges”, in Proceedings of the 12th EC GI&GIS Workshop. Innsbruck (Austria).

Gould, M., 2005. Meta-Findability: Part 1. GEO:connexion, Magazine available at http://www.geoconnexion.com/uploads/meta_intv5i7.pdf

GSDI. Global Spatial Data Infrastructure association. www.gsdi.org (accessed 3 May 2007).

Granell, C, Gould, M, Manso, M.A, and Bernabé, M.A. 2008. Spatial Data Infrastructures. In H. Karimi (Ed.): Handbook of Research on Geoinformatics, Hershey Pennsylvania: Idea-Group.

gvSIG. Proyecto gvSIG. http://www.gvsig.gva.es/

Manso, M.A., Nogueras, J, Zarazaga, J, Bernabé, M.A. 2004. “Automatic Metadata Extraction from Geographic Information”, in Proceedings 7 AGILE Conference on Geographic Information Science, University of Crete, Greece, pp. 379-385.

Nogueras, J, Zarazaga, F.J, and Muro, P. 2005. “Geographic information metadata for spatial data infrastructures”, Springer-Verlag.