

# Semi-automatic Metadata Extraction from Imagery and Cartographic data

Laura Díaz, Cristian Martín, Michael Gould,  
Carlos Granell

Centre for Interactiva Visualization (CeVI)  
Universitat Jaume I  
Castellón, Spain  
laura.diaz@uji.es

Miguel Angel Manso

Department of Topographic and Cartographic Engineering  
Universidad Politécnica de Madrid  
Madrid, Spain  
m.manso@upm.es

**Abstract**— Metadata are necessary to allow discovery and description of data and service resources within a Spatial Data Infrastructure, however current manual metadata editing workflows are tedious and under-utilized. We discuss on-going developments for semi-automatic metadata extraction from well-known imagery and cartographic data sources, being implemented within an open source software project in Spain. Internal metadata are collected automatically and the user can then choose to add external metadata, and to publish the final metadata record to catalogues. The next step will be to extract implicit metadata using Google-like methods.

**Keywords**— metadata, information retrieval, open source, Spatial Data Infrastructures

## I. INTRODUCTION

Metadata is a key element for allowing optimal data fusion and discovery, and for Spatial Data Infrastructures (SDI) [1, 2] to operate properly. Most Earth observation (EO) image processing software is equipped to read and exploit the image header file, a common type of internal metadata, to learn more about the image characteristics (size, coordinate system, pixel resolution, etc.) and, thus, properly visualize and process it. However, these internal metadata are normally not compete enough to assist the human user in judging whether the image is useful or not, covers the proper geographic area (bounding box is not enough), has important cloud cover, etc. It is normally the user who must manually create these external metadata, which are necessary to be able to publish, search for, and facilitate access to that data product in an SDI. Creating the documentation describing who created the data, where they can be found, what geographic places they cover, their general description or abstract, etc., is a laborious task, in part because users often possess data created by other parties and, so, it can be difficult to locate the original or any knowledgeable sources for some key metadata elements.

This documentation process should be automated to the extent possible, given that informatics technology has greatly improved since the early days of the digital libraries that gave birth to the current manual metadata creation methodology.

This process is currently undertaken using simple text editors and outside of the GIS or image processing workflow. The metadata problem is greater still among users of digital cartographic (vector) data, because GIS software managing these data has less of a tradition of including metadata in a header-like file (an exception being so-called world files, containing coordinate reference information) and is often even less capable than image processing software with regard to metadata extraction and exploitation. 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 [3] and GMES [4] are promoting heterogeneity and interoperability, making the availability of open source solutions all the more attractive.

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. This is the direction we have chosen to follow [5].

## II. METADATA CREATION PLATFORM

A large migration project from proprietary to free software, initiated in 2004 by the Valencia regional government (Generalitat Valenciana), has produced a client software product called gvSIG ([www.gvsig.gva.es](http://www.gvsig.gva.es)) [6]. What began as a simple, Java-based GIS client quickly evolved to become, at its version 1.0, a full-function 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 to semi-automatically extract metadata from well-known geodata formats (GeoTIFF, Shapefile, etc.), that is, at the dataset (layer) level. The idea, as stated previously, is that GIS/SDI users are given the ability to document 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. The gvSIG user is given a new menu item, which allows him or her to view the metadata describing the selected geodata layer, to edit that metadata, and to finally publish the metadata, if so desired.

### III. METADATA EXTRACTION AND PUBLICATION

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, for the moment called MetaData Markup Language, MDML.

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://eu-geoportal.jrc.it/gos>).

At the moment a well-know 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 [7] 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. [8]. 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 Catalog 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. If

not valid the manager informs the user of those metadata elements missing or incorrect. The objective here is to not overly restrain the user, but on the other hand also to not allow publication of metadata records that will not be useful according to catalogue and future search requirements.

### IV. CONCLUSIONS

We demonstrate one possible and practical implementation of the concept of semi-automatic metadata extraction and management, to assist users who create or edit imagery or cartographic data and then wish to publish these data in an SDI environment. 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. Future work will include import functions for geodata already possessing metadata created elsewhere, and intuitive extraction of metadata following intrinsic (or context-based) clues within the data resource itself: the so-called Google methodology. This will include tasks such as automated extraction, using deductive methods, of information sufficient to generate a credible and useful free-text abstract describing the dataset.

### ACKNOWLEDGMENT

This research has been partially supported by the regional ministry Conselleria de Infraestructuras y Transporte (Generalitat Valenciana).

### REFERENCES

- [1] GSDI. Global Spatial Data Infrastructure association. [www.gsdi.org](http://www.gsdi.org) (accessed 3 May 2007).
- [2] C. Granell, M. Gould, M.A. Manso, M.A. Bernabé. Spatial Data Infrastructures. In H. Karimi (Ed.): Handbook of Research on Geoinformatics, Hershey Pennsylvania: Idea-Group, 2007 (in press).
- [3] 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/> (accessed 3 May 2007).
- [4] European Commission Directorate General Enterprise, GMES: Global Monitoring for Environment and Safety. [www.gmes.info](http://www.gmes.info) (accessed 3 May 2007).
- [5] M. Gould, J. Rocha, S. Nativi, J. Nogueras, M. Manso. "Near-term metadata challenges", in Proceedings of the 12th EC GI&GIS Workshop. Innsbruck (Austria), June 2006. <http://www.ec-gis.org/Workshops/12ec-gis/programme.cfm> (accessed 3 May 2007)
- [6] M. Gould, C. Granell, M.A. Esbrí, G. Carrión, "The role of free software thick clients in SDI: Case of gvSIG", in Proceedings of the 12th EC GI&GIS Workshop. Innsbruck (Austria), June 2006. <http://www.ec-gis.org/Workshops/12ec-gis/programme.cfm> (accessed 3 May 2007).
- [7] GDAL: Geographic Data Abstraction Library, <http://www.gdal.org/> (accessed 3 May 2007).
- [8] M.A. Manso, J. Nogueras, J. Zarazaga, M.A. Bernabé, "Automatic Metadata Extraction from Geographic Information", in Proceedings 7 AGILE Conference on Geographic Information Science, University of Crete, Greece, pp. 379-385.