Automatic Metadata Extraction from Geographic Information

M.A. Manso[1], J.Nogueras-Iso[2], M.A. Bernabé[1], F.J.Zarazaga-Soria[2]
[1] Department of Topography and Cartographic Engineering
Polytechnic University of Madrid
Campus Sur, Carretera de Valencia, km 7. 28031- Madrid (Spain)
m.manso@euitto.upm.es, mab@euitto.upm.es
[2] Department of Computer Science and Systems Engineering
University of Zaragoza
María de Luna 1, 50018 Zaragoza (Spain)
jnog@unizar.es, javy@unizar.es

ABSTRACT

One of the biggest problems for implementation of a Spatial Data Infrastructure is the creation of metadata because it requires a high knowledge of sciences related to cartography and a good infrastructure of tools that allow to get information from the original Geographical Information and which carry out transformations or conversions of coordinates, transcription of multiple information, etc.. This fact has motivated a detailed study of file formats of geographical information. The objective was to identify the information which can be acquired from its own files and to study how they are related with the metadata standard of geographical information. The study has shown different techniques used by enterprises to store meta information in form of headers, directories and labels. Homogeneous groups of information which can be retrieved from the different categories of formats have been identified. This study has provided a high range of conclusions and perspectives which should be guidelines for future work.

KEYWORDS: *Geographic Metadata, file formats for geographic information, extraction of metadata*

INTRODUCTION

Metadata are "data about data", that is to say, they describe the content, quality, condition, and other characteristics of data in order to help a person to locate and understand data. The creation of metadata has three major objectives [FGDC 2000]. The first one is to organize and maintain an organization's investment in data. As personnel change or time passes, information about an organization's data will be lost. And later workers may have little understanding of the content and uses for a digital database and may find that they cannot trust results generated from these data. Complete metadata descriptions of the content and accuracy of a geospatial data set will encourage appropriate use of the data. Such descriptions also may provide some protection for the producing organization if conflicts arise over the misuse of data. The second objective is to provide information to data catalogues and clearinghouses. Applications of geographic information systems often require the integration of data from different thematic sources. And few organizations can afford the creation of all data they need. Furthermore, data created by an organization also may be useful to others. By making metadata available through data catalogues and clearinghouses, organizations can find data to use, partners to share data collection and maintenance efforts, and customers for their data. Finally, third objective of metadata is to provide information to aid data transfer. Metadata should accompany the transfer of a record. The metadata aids the organization receiving the data process and interpret data, incorporate data into its holdings, and update internal catalogues describing its data holdings.

The first formal definition of the term "National Spatial Data Infrastructure" was formulated in the US and published in the Federal Register on April 13, 1994 [Federal Register 1994]. It states: "National

Spatial Data Infrastructure (NSDI) means the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data". Maybe the principal problem for the implementacion of a Spatial Data Infrastructure is the absence of catalogs with appropriate and well-defined metadata. Taking into account that the ISO/DIS 19115 standard defines more than 300 metadata elements, it can be assured that the creation of appropriate content for all the different metadata records is a hard and arduous process. Therefore it might result very interesting to have a tool that automatically extracts metadata from data sources. Besides saving time in the cataloguing process, it prevents users from making frequent typing mistakes.

This paper presents a work done in order to facilitate the automatic metadata extraction from a great variety of geographic-data formats used for the storage of geographic information. The rest of this paper is structured as follows: Next section makes a revision of the most common storage file format for Geographic Information structured by the type of its content. For each category the main properties that may be used for the registering process of metadata are described and some tables are presented that summarize metadata extraction properties for each file format. Finally, this paper ends with conclusions and future work section.

REVISION OF THE FORMATS OF COMMONLY USED FILES TO STORE GEOGRAPHIC INFORMATION

Geographic information can be stored in several distinct manners depending on its proper nature. Thus we can obtain raster images, vector files, regular grids, tabular data or data base structure as a means of storage. Each country adopts a set of file formats as standard of interchange of geographic data. In some cases these file formats are patented by enterprises of the same sector. Each country or region organizes its own geographical information following its own criteria. In some cases continues information format is used and in other cases information is distributed in tiles. This chapter will intend to provide a classification of different files format based on the nature of stored data. The next step will be the enumeration of metadata that may be stored and retrieve from the given information. The main objective is to study what type of metadata of the standard ISO 19115 may be obtained from its original file storage. As a summary of this section several charts are presented which synthesize the above-mentioned.

Raster files of general purpose

Raster file formats of general purpose are: BMP (bitmap format), PNG (a portable network graphic format), RAS (raster file format used by GRASS GIS), TIFF (Tagged Image File Format used by Adobe and others), JPEG (Joint Photographic Experts Group), GIF (Graphic Interchange Format), IFF (Interchange File Format) and PCX.

After studying the technical documents that describe these raster file formats it has been shown that they have the following characteristics in common: Dimensions of the image (width and height) are expressed in number of pixels, the number of bands and components, the number of bits applied to represent each pixel and the type of compression used. It has been observed that most of the applications that make use of this type of formats to store geographical information use a simple technique that allows to have a coordinate system to handle images. This technique is based upon the use of an auxiliary text file, commonly indexed as World File; This file contains all the necessary parameters to carry out a transformation (rotation, translation and size transformation) through: The coordinates of the point of the top left image corner (minimum X and maximum Y or East Longitude and North Latitude), the size of the point according to its axis and as well as two additional parameters that allow to define the turns on the image. This World File shares the name of the file that contains the image and adopts one of the following extension files: wld, tfw, jpw, etc… The main inconvenience of this technique is the missing information about of space reference system (datum and projection) to which the stated coordinates make reference. In most of the cases coordinates are flat or projected coordinates. In spite of the above-mentioned, it allows a better knowledge of geographical information represented in form of metadata. In

continuation the information fields are enumerated which can be obtained from the enumerated sources: number of columns, number of rows, number of bits by pixel, number of bands, type of compression, horizontal units and the coordinates of the BoundBox. For some file formats as the PNG and the TIFF exists the following additional information: author, content description, creation date and a source list.

Raster file formats used to store digital aerial photographs, orthogonal images & scanned cartography

Raster file formats used to store digital aerial photographs, ortho images, scanned images, etc… with high information size: GeoTIFF (Tagged Image File Format which comprises information concerning the special reference system), MrSID (file format with compression based on wavelets by LizarTech Inc.), ECW (file format with wavelet compression technology by Ermapper), JPEG2000 (standardized compress file format based on wavelet technician ISO 15444), GeoJP2 (JPEG2000 file format with information related to the spatial reference system stored as GeoTIFF by Mapping Science), INGR (Intergraph raster file format) and NITF (North American standard of transfer of images of the National Imagery and Mapping Agency: NIMA).

The first difficulty with which we were confronted when studying the raster file formats used to store great images (having as its origin aerial photographs transformed to orthogonal projection, etc..) is the absence of a public document who defines the file format. This had happened with the file formats: MrSID and ECW. These difficulties have been palliate partially with existing software tools [SiDV 2.1], [ECWHE 2.5] that extract information from the headers file and this gives a hint about what type of metadata is stored. The common information that can obtained from the analyzed formats in this work (GTIFF, MrSID, ECW, JPEG2000, GeoJP2, INGR and NITF) and which is useful for the metadata content are as follows: image dimensions (width and heigth) expressed in number of pixels, number of bands, number of bits by pixel, type of compression, distance units of pixel in each axis, measuring units, the maximum and minimum coordinates (that define the BoundBox) in a system of projection defined by the datum, ellipsoid and the projection. In some cases the codification of the information associated to the space reference system is standardized as it happens with GTIFF format, in other cases particular codifications are used. For some formats a large number of metadata may be retreived like: creation date of content, quality of compression, statement about data source, use and/or access restrictions to information or finally more specific parameters of the spatial reference system (zone, length of the greater half axis of the ellipsoid, flattening coefficient value, longitude and latitude of origin, etc..).

It is important to mention that the work of standardization was made applying the GeoTIFF format that is characterized through a set of marks that allow manage the parameters related to space reference system within TIFF files. The same technique is used by GeoJP2 file format. GeoJP2 is one specialization of the JPEG2000 format that includes the information about spatial reference system in form of an image of a single pixel in a file block of the type UUID (Universal Unique Identifier). Also it should be emphasized upon the facility of the JPEG2000 format to include a block of textual information, in form of a file of the type XML, within the JP2 file containing metadata. An international standard exists that define types and the organization of metadata related to content and the intellectual property that applies to the archives with the compression JPEG2000 (ISO 15444-2).

Digital terrain models stored as raster grids

Digital terrain models stored as raster grids: ADF (ArcInfo Data Format GRD of Esri), GRD (Ascii Grid format of Esri), GRD (Surfer grid file format of GoldenSoft), DEM (Digital Elevation Model of USGS), DEM (Digital Elevation Model used for MicroDEM: Professor Peter Guth), DTE (Digital Terrain Elevation of Socet Set), DT0 (Digital Terrain models of the American Department of Defence), HGT (Digital Elevation Models of the project Shuttle Radar Topography Mission of NASA), BIl, BIP and BSQ (MapInfo raster interchange file formats).

After searching for and examine existing documentation relative to the formats of regular grid used to store the elevations of terrain it can be concluded that there exist some multiple formats, some of them standardized and others defined by companies. The files formats that have been analyzed are: BIL, BIP

and BSQ of MapInfo, Gtopo30, Raster export format from Erdas Imagine, HGT of the mission SRTM, ADF and GRD of Esri, Grid of Surfer, DEM of the USGS, DEM of MicroDEM, DTED, DOQ2 and DTE of Socet Set. It has been verified that there exists a great similarity between all the enumerated formats concerning metadata. A big difference might arise when the format does not store information of spatial reference system (datum, ellipsoid, projection, zone, parameters). In some cases this information is implicit to the format (GTOPO30, HGT). For the rest of the cases (DTE of Socet Set and Grid of Surfer) the longitudes and latitudes of the bounding box which limit the space elevations model cannot be computed. The information that have all formats in common is as follows: Dimensions of the grid (number of rows and columns), the increase of coordinates with each pixel on each axis, horizontal units, maximum and minimum levels in the elevation model, the projection system, datum, ellipsoid and projection parameters.

Raster files formats used by remote sensing software to store satellite images

Raster file formats used by remote sensing software to store satellite images: IMG (Erdas Image file format), PIX (PCI and Geomatics image data base file format), ERS (image file format of Ermapper), IMG (image file format of Idrisi), NOAAL1B (file format to distribute low resolution images of the AVHRR sensor of NOAA Satellites) and the fast distribution file format EOSAT (Earth Observatory Satellite) F-EOSAT (fast distribution file format for LandSat 7 and IRS satellite).

It has been tried to analyse the huge number of file formats used by the software of remote sensing. After a first evaluation process and before starting with the analyse of the documentation enough information was collected to determine what existing information within the storage format may be used as metadata. This analysis has been made with the following formats: IMG, PCIDISK, ERS, IDRISI, NOAAL1B and F-EOSAT (Fast Earth Observatory Satellite). All of these formats have in common the property that they store abundant important information concerning metadata but part of them do not enter the standard ISO 19155. It is important to assure the appearance in the second norm (ISO 19115-2) that tries to gather the metadata necessary to describe the grid of information like the information from satellite photos in remote sensing technology. The metadata that may be read from the image files are: number of bands of the scene, dimensions of them (width and height) expressed in number of pixels, horizontal units, pixel size, projection system, datum, bounding box coordinates that limit the scene (although in some cases the coordinates are provided of four corners and one of the centre), data type used to store pixels, acquisition date, satellite platform, reception station that applies the pre-process, pre-process level, statistical parameters of the digital values of pixels, parameters of the acquisition system position and angles and other parameters that can be difficult to insert into metadata.

Vector file formats

Most extended vector file formats are: DGN (Design file format used by Bentley and Intergraph), CSF (Coordinate System File of Intergraph), DWG and DXF(draw file format applied with Autocad: Autodesk), ADF and E00 (Coverage file format of ArcInfo: Esri), SHP (Shape file format of the entities of ArcView:ESRI), MIT and DAT (map file formats of MapInfo), VEC (vector file format of Idrisi), BIN (binary file format of Digi: Digi21).

It has been intended to collect documentation of vector file formats that are commonly used to store geographic information generated by the software of computer aided design (CAD) and could have been part of the Geographic Information System (GIS) projects. In some cases there is not an explicit documentation of the file's origin, in other cases exists certain information which those files do not describe. The set of files formats analysed by the study are: E00, SHP and ADF of Esri, DGN and CSF of Intergraph, DXF and DWG of Autodesk, MIF and TAB of MapInfo, IMG of Idrisi, BIN of Digi21 and GML of OpenGis. Like it happened with images files formats, many of these formats do not have fields of information of headings or structures of archives to store the information concerning spatial reference system in which the coordinates of the entities are described. This is the case of the formats: DGN, DXF, DWG, BIN. Nevertheless the format DGN can be accompanied of a Coordinate System File (CSF), even it can incorporate a information block that contains this same information within its own file information

Format	W / H	BoundBox	pixel resolution	Bits / pixel	Bands / dimensions	Max, min statistical	Other stadistics	Horizontal Units	Projection	Datum	Ellipsoid	Other
Gif	X	X(1)	X(1)	X	X							
Png	X	X(1)	X(1)	X	X							
Jpg	X	X(1)	X(1)	X	X							
Iff	X	X(1)	X(1)	X	X							
Tiff	X	X(1)	X(1)	X	X							
GeoTiff	X	X	X	X	X			X	X	X	X	
Bmp	X	X(1)	X(1)	X	X							
Pcx	X	X(1)	X(1)	X	X							
Psd	X	X(1)	X(1)	X	X							
Ras	X	X(1)	X(1)	X	X							
Sid	X	X(1) (2)	X(1) (2)	X	X			X(2)	X(2)			
Ecw	X	X(1) (2)	X(1) (2)	X	X			X(2)	X(2)	X(2)		
JP2000	X	X(1)	X(1)	X	X							
GeoJP2(JP2k)	X	X	X	X	X			X	X	X	X	
DOQ2 (USGS)	X	X	X	X	X			X	X	X		
DTED	X	X	X	X	X			X	X	X	X	
DTE-Socet set	X	X	X	X	X			X				
GRD ESRI	X	X	X			X(3)		X	X	X	X	
GRD surfer	X	X	X			X						
DEM -USGS	X	X	X	X		X		X	X	X		
MicroDEM	X	X	X	X	X	X		X	X	X		
E00 grd	X	X	X(4)		X	X(4)	X(4)	X(4)	X(4)	X(4)	X(4)	
NTIF	X	X	X	X	X			X	X			X
ADF grd	X	X	X(4)		X	X(4)						
Lan erdas	X	X	X	X	X							
IMG erdas	X	X	X	X	X	X						
PIX	X	X	X		X							
ERS	X	X	X	X	X	X						
DOC (Idrisi)	X	X	X	X		X		X	X	X	X	

Table 1: Summery of the information that may be obtained from raster files.

[1] In case of the existence of a world file.
[2] If version file contains this information.
[3] If all the data file has been read and its values may be computed.
[4] If all the file sections are presented

within its own file DGN (element type 56). The format of the file CSF and field 56 (used by MGE), is not public although there exist tools that allow to read the content of them such us [DefCSF] or that save this information in a text file form such us [CSF2TXT].

The SHP format uses a special semantics to express the spatial reference system in the PRJ file with a data structure known us Well-Known Text (WKT) [WKT 2001]. Common information that can be obtained from the vector files with Geographic Information content are: the maximum and minimum

coordinates of the bound box rectangle that limits entities, the number of existing entities of each type (points, shapes, poly-lines, circles, labels, etc), the name of the layers of drawing associated with the entity type, spatial reference system, datum, ellipsoid and projection parameters. In some cases it may be possible to obtain additional information related to the processes applied to information, date of such process, information sources in usage, etc…

SUMMARY

Tables 1 and 2 represent a summary of different types of geographic information. One is a summary of raster formats. The other one describes the vector files formats. These tables exhibit information that may be obtained from the files and that may be directly integrated in metadata processing concerning the ISO 19115 standard.

Format	BndBox	Nº of layers	Name of layers	Nº of diferent features	Name & number of features	Horizontal Units	Projection	Date	Ellipsoid
E00 arc	X					X	X	X	X
ADF arc	X			X	X	X	X	X	X
DGN	X	X		X	X				
DXF	X	X	X	X	X				
SHP	X	X	X	X	X				
MIF	X			X	X	X[(1)]	X[(1)]		
TAB	X			X	X	X	X	X	X
DWG	X	X	X	X	X				
VEC (idrisi)	X			X	X	X	X		
BIN	X	X[(5)]	X[(5)]	X	X				

Table 2: Summary of information that may be obtained from vector files

[1] Equivalent semantics to the names of the layers

CONCLUSION

One of the biggest problems for implementation of a Spatial Data Infrastructure is the creation of metadata because it requires a high knowledge of sciences related to cartography and a good infrastructure of tools that allow to get information from the original Geographical Information and which carry out transformations or conversions of coordinates, transcription of multiple information, etc.. This fact has motivated a detailed study of file formats of geographical information in order to identify the information which can be acquired from its own files and to study how they are related with the metadata standard of geographical information. The processes of information collection and applied analysis on those more than 45 formats of files have provided a better insight into techniques used for the same ones to structure the information (directory, labels, heading, attached files, etc.). The formats of files studied have been classified in function of the type of information which they usually store. Following this guideline common information has been distilled which are stored in headings, labels or attached files. In an indirect way there has been described a pseudo metadata pattern for the different file types.

Semantic inconsistencies among the different formats of files have been detected in the representation of the spatial reference system using different terminologies and storage forms. Even a total absence of this type of information was detected in several formats. The harmonization of the semantics associated with the spatial reference system is proposed as a interesting direction of investigation in order to express this information obtained from each format of analyzed file in a standardized way.

There has been detected abundant descriptive information of some formats of image files used for remote science which are not listed in the standard ISO 19115. As a future goal is proposed to establish a semantic relationships between these fields and the ones of new metadata which are defined in the working paper for images and remote science (ISO 19115-2.).

It has been discovered the existence of other metadata standards to describe the content and the intellectual property associated to the images JPEG2000. A creation of a crosswalk of metadata is proposed that might appear in the standard ISO 15444-2 up to the standard ISO 19115.

Finally, another work that has been launched is the analysis of the existent computer tools to study its capacity to obtain an automatic extraction of geographic information which should be structured afterwards in a file format including metadata.

ACKNOWLEDGEMENTS

The basic technology of this work has been supported in part by the Ministry of Science and Technology through the project TIC2000-1568-C03 of the National Plan of Scientific Investigations, Technological Innovation and FIT-150500-2003-519 from the National Plan for Information Society.

REFERENCES

CSF2TXT. Coordinate System File dump to text tool from Intergraph.

DefCSF. CSF & DGN Coordinate file editor from Intergraph.

ECWHE v2.5. ECW Header Editor v2.5.1. Ermapper Corp. www.ermapper.com

Federal Register 1994. U.S. Federal Register, "Executive Order 12906. Coordinating Geographic Data Acquisition and Access: the National Spatial Data Infrastructure (U.S.)", The April 13,1994, Edition of the Federal Register, vol 59, number 71, pages 17671-17674.

FGDC (2000). Content Standard for Digital Geospatial Metadata Workbook, version 2.0. Federal Geographic Data Committee (USA), 2000.

SiDV, 2.1. MrSID GeoViewer 2.1 Lizardtech Corp. www.lizardtech.com

WKT, 2001. Coordinate Transformation Service, Well-Known Text Format. 2001 http://www.opengis.org/