Image Metadata: Compiled proposal and implementation

Alaitz Zabala
Centre de Recerca Ecològica i Aplicacions Forestals, CREAF, Bellaterra, Barcelona, Spain

Xavier Pons
Departament de Geografia Universitat Autònoma de Barcelona & Centre de Recerca Ecològica i Aplicacions Forestals, CREAF, Bellaterra, Barcelona, Spain

ABSTRACT: Metadata about remote sensing data have special needs to properly document topics that are not needed in other type of data (such as platform and mission information, multi-band images,…) or that need special consideration when referencing to remote sensing data (such as reference system, spatial extension, bounding quadrangle, pixel size, resolution,...). A metadata model has been developed to consider those aspects. A first look at the model as well as software implementation considerations are explained in this paper.

1 INTRODUCTION

Metadata describe the content, quality, condition, and other characteristics of data. Although its obvious interest, it is not under the last years that important efforts have been paid to this topic applied to geoinformation disciplines. Among these efforts, we should consider there are the *Content Standard for Digital Geospatial Metadata* (FGDC 1998), the European pre-standard *Geographic Information - Data description - Metadata* (CEN 1998), and the contributions of the Open GIS consortium (*OpenGIS Abstract Specification - Topic 11: Metadata*). On the other hand, the International Organization for Standardization Technical Committee 211 has published a Draft International Standard (DIS), *Geographic information – Metadata*, on September 2001 (ISO 2001). The Final Draft International Standard was supposed to be approved in April 2002 and the International Standard should be published in July 2002.

When considering these valuable approaches, it can be noted that the field of Remote Sensing has not been considered in depth. As a way to alleviate this problem, the FGDC is developing a standard on Remote Sensing Metadata (FGDC 2000) actually at Step 9, Respond to Public Comments of the FGDC Standards Process, but up to date we do not know a similar initiative from the European or other context.

Besides the specific modules used in each software package, there are few software implementations to properly document Metadata (either about images or not). Even the specific modules to document and modify metadata given by some important remote sensing software packages are very simple.

Some of the packages revised by the authors have just a two-column list to document metadata entry names and their values, and just a few of these names are explicitly recognized as standardized metadata entries. Even other software still has no tools to document metadata.

This lack of appropriate software is probably due to the relatively recent interest on the topic, but also due to the difficulty on finding an accepted standard and a consensus between the large amount of entries on the proposed standards and pre-standards. It is necessary to reach a consensus point between a simplified approach (operative, but in the long run insufficient) and the complex and vast existent proposals (often impossible to be compliant with). An optimal solution should be easy to use but extensible to the most complex needs.

The standards define different sections and entries that have to be filled to properly document a geographic data set. The entries defined in the standards may be mandatory, mandatory if applicable (if the data set exhibits the defined characteristic) or optional. FGDC and ISO standards also have an extension mechanism to permit users defining a set of metadata entities and/or elements used by a specific discipline or organization.

FGDC has developed a *Extensions for Remote Sensing Metadata Standard* "to provide a common terminology and a set of definitions for documenting geospatial data obtained by remote sensing" (FGDC, 1998). The FGDC standard complies with ISO extension methodology and, as they explain at 'Related Standards' section ('Introductory Material' chapter), the "Remote Sensing Extensions have been con-

structed to be compatible with that (ISO extension) methodology" (FGDC, 1998).

The aim of the study explained in this paper is to elaborate a metadata model, result of a compilation of the existing standards and some extra adds from our point of view, and a software implementation of that model.

This paper explains some of the considerations that have been taken into account during the definition of the metadata model and in the software implementation.

The metadata model is developed to be compliant with the cited standards but also to be a unique model to properly document metadata in each context. Software implementation allows the user to document its data sets with this metadata model and will offer the possibility to export the metadata model (in REL format file) to ISO or FGDC model (using HTML or XML format).

2 COMPILED PROPOSAL: SOME CONSIDERATIONS

At the present moment, not all the sections and entries of the related standards have been included in the model. Sections about identification, reference system, spatial data and data quality are considered up to date and other aspects are in progress.

The major aspect considered during the definition of the model is to develop easy to use but extensible model. It is known that the communities that creates and uses geospatial data in each context usually do not need all the complexities defined in the standards. One example of this situation can be found in 2.2 *Reference System Information*.

Another important aspect considered in the model is the normalization of the possible contents of some entries that in a first implementation should be treated as a free text entry. (e.g. name of quality parameters and measures: 2.4.1 *Quality parameters*, name of units,…).

2.1 Multi-band and related topics

The FGDC standard defines a data set as a "collection of related data". So it can be considered that different bands of a multi-spectral or hyper-spectral sensor form the same data set. In a similar way, the different images representing height, slope, aspect,... in a digital terrain model (in a GIS context) could be considered related data so they could form another multi-band data set.

It should be possible, of course, to define metadata of each band separately. Nevertheless it is usually more interesting to document all the bands together because many of the entries of the different bands have the same value. So documenting all the dataset as a whole would be more consistent and eas-

ier to revise and update than if all the bands are documented independently. It may be considered that you have different information for each pixel. For example you may store for the same location, the radiance in red, green and blue wavelength or the height, slope and aspect of the same pixel.

On the other hand, this aggregation of metadata must be carefully treated in some aspects to avoid improperly documentation of data or inconsistencies between real and documented metadata. It will be discussed in 3.1 *Multi-band and related topics implementation*.

Some restrictions are necessary to implement this point of view. To create one multi-band data set it is necessary that all the bands have the same reference system to create one multi-band data set. It is strongly recommended (but not necessary) that all the bands refer to the same spatial extension (but not the same temporal extension, because even may be interesting to group different date images from the same area). That is because it may be considered, as said before, that different band values are referred to the same spatial location.

However this latest recommendation is not always followed because of the same process of capturing the images. One example, bands derived from some airborne digital cameras are spatially out of phase (in the track direction) because of the movement of the airplane during the caption and the delay among the capture of the different spectral frames. Another case is Landsat 5 TM images that originally have different resolution (and different pixel size) in thermal band. If the pixel is not resized, the thermal band of the multi-band would have the same spatial extension but different row and column count an different pixel size than the others (change on pixel size do not change resolution of the band which will always be 120m for the thermal band).

It should be considered three types of entries: those which apply just to the dataset as a whole, those with a general value that just applies to bands without any specific value and those which apply just to each band (without any general value because it has no sense). One example of an entry that just applies to the dataset is the Dataset Title. Pixel size, bounding quadrangle and column and row count are examples of those entries with a general value (the value of the major number of bands) that applies to bands without a specific value in this entry. Documenting in that way is useful because you have to store each value entry just once and you have to document band values only if they are different from the general value. Finally, such entries as band name, value treatment, minimum and maximum value,... do not have any general value because they apply to each band individually.

2.2 *Reference System Information*

Standard sections to document reference system are very vast and complex. There are lots of entries that should be filled to document in depth these metadata elements. The names and number of entries varies enormously depending on which reference system has to be documented (direct or indirect system, geographic, planar or local,...).

Moreover, users usually only need a few of this complex possibilities. This is why should be considered a simple way to document reference system that allows users to simply choose one of the predefined reference systems (defined by an specific projection, datum, ellipsoid, etc). It should exist the possibility of defining new reference systems (with other parameters).

Our implemented model permit define any reference system by choosing de proper projection system, datum, ellipsoid, etc. Each particular combination of these parameters is known by an unique identifier defined in a table and related to one particular datum, projection, ellipsoid, etc (defined in other tables). The user just has to select one of those pre-defined reference systems (which are a list of those useful in each context, e.g. UTM-31N with European Datum 1950, International 1924 ellipsoid, etc).

On the other hand, every special need of the user can be implemented just defining in those tables a new identification to the reference system needed and all its parameters. For example reference system "UTM-31N WGS84" should be related to "WGS 1984 Datum" and "WGS 1984 ellipsoid" (this one is also provided as a defined one).

A caption of the software implementation window to document those entries is shown at Fig. 1.

2.3 *Geometry*

After geometric corrections needed to adapt raw data to a specific spatial reference system, bounding quadrangle of the image usually has some decimal figures in each coordinate. These supposed precision it is not real because it is the result of the mathematic process of adjusting the lineal regression model used to geometrically correct the image (usually the actual resolution is related to the original pixel size, that is greater than that precision).

This is why sometimes should be interesting to slightly modify those coordinates to reach an agreement between bounding quadrangle coordinates of different bands or images that really have the same resolution and geometric spatial position but different coordinate decimals due to geometric correction process.

Related to this aspect is very important the interdependency between pixel size (different from resolution), bounding quadrangle coordinates and column and row count. To avoid inconsistencies in those values, the possibility of introducing incoherent values should be protected (see 3.2 *Geometry implementation* to see how software implementation consider that point). In fact, pixel size is never stored in our metadata model (but it is really stored in export formats) because it can be calculated from other present values.

2.4 *Quality and lineage*

Most standards define a Data Quality section, which basically include two conceptual elements.

First of all the way to document the quality of a metadata element (simple or compound) is provided by defining quality parameters which inform about the type of test applied to the data set (or to one metadata element) to document its quality (giving quality measures).

On the other hand, lineage describes information about the events or source data used in constructing the data set.

2.4.1 *Quality parameters*

A tree structure is defined to document measures that define the quality of an element (for example the positional accuracy of a dataset, or the semantic accuracy of a classified image).

Each quality parameter should have at least one quality indicator that, at the same time, should have at least one quality measure. Each parameter is related to one aspect of the quality of the dataset (positional accuracy, temporal accuracy, completeness, semantic accuracy, etc). For example a "positional accuracy" parameter may have a "relative horizontal accuracy" indicator and different measures such as RMS, RMS on x, RMS on y, etc.

This tree list is very useful to aggregate different measures of quality related to the same aspect (because they will be depending on the same quality parameter). This structure allows avoiding an unordered two-column list with quality measures and its values.

Related to this topic may be considered the necessity that metadata entries could have defined values that everybody understands in the same way. These definitions are useful to avoid people defining new names for the same concepts. Standards (and specially ISO/DIS-19115 with its code-lists) usually define the domain for some entries (although other entries are defined as "free text"). Those definitions are universal and should not be ignored. Nevertheless, sometimes more local concepts may be defined to promote understanding among users in the same context. Quality parameters, indicators and measures are a good example of this situation.

To see how the software implementation considers these aspects refer to 3.3 *Quality and lineage implementation*.
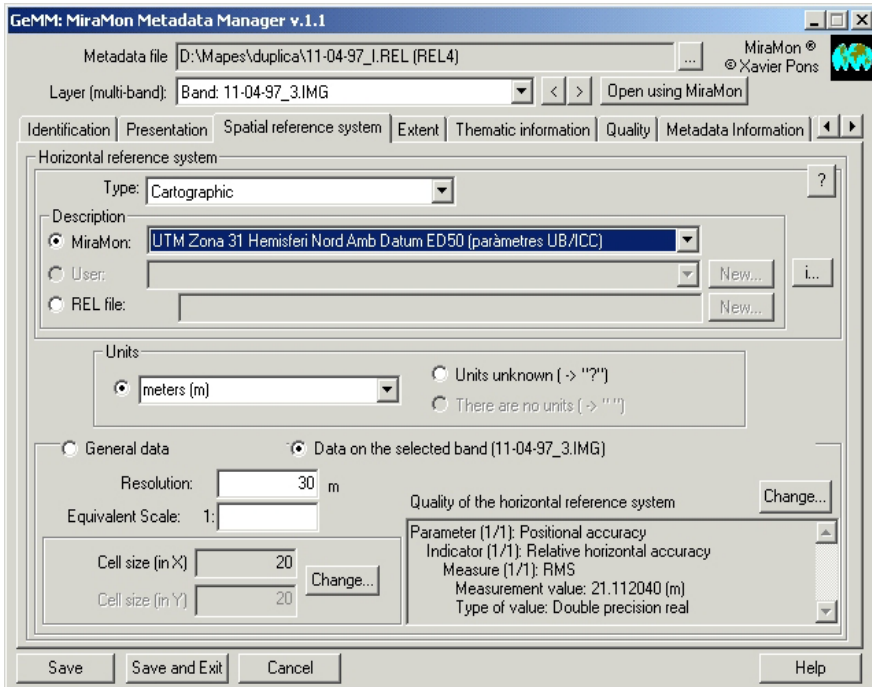
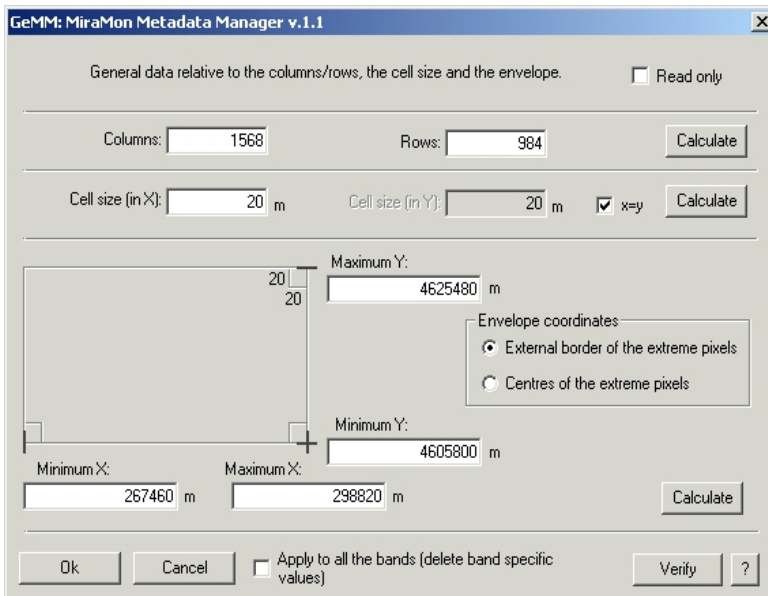Figure 1. Software implementation: spatial reference system metadata

Figure 2. Software implementation: column and row ow count, pixel size and bounding quadrangle

Zabala , A., Pons, X.

2.4.2 Lineage

To properly understand the values stored in the image (or in the data set), it is necessary to document all the processes that have been executed over the data set and the different sources used to create it.

It should be noted that some processes made over the data set might need other data sets to be done. It is necessary to document the lineage of data set sources used in a process. That is especially important in a Remote Sensing context where multi-source processes are usually made over the datasets.

This is why every source defined (as a source of the dataset or as a needed source of a process) may have a related lineage section defining processes done to the source till de moment it is used to produce de dataset or to taking part in one process made over the dataset.

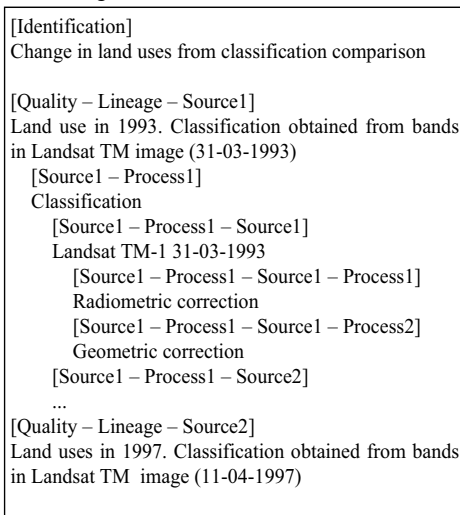A conceptual model of that implementation can be found in Fig. 3.

```
[Identification]
Change in land uses from classification comparison

[Quality – Lineage – Source1]
Land use in 1993. Classification obtained from bands
in Landsat TM image (31-03-1993)
   [Source1 – Process1]
   Classification
     [Source1 – Process1 – Source1]
     Landsat TM-1 31-03-1993
       [Source1 – Process1 – Source1 – Process1]
       Radiometric correction
       [Source1 – Process1 – Source1 – Process2]
       Geometric correction
     [Source1 – Process1 – Source2]
     ...
[Quality – Lineage – Source2]
Land uses in 1997. Classification obtained from bands
in Landsat TM  image (11-04-1997)
```

Figure 3. Lineage: processes and sources (conceptual model)

3   IMPLEMENTATION

Simplicity and usefulness has been the base objective in software implementation. Help is offered at different levels (software operation, section information and entry information) to facilitate metadata generation and maintenance by the user.

Furthermore other ideas have been considered during the implementation of the metadata model as discussed below.

3.1   Multi-band and related topics implementation

As commented in previous points, it is necessary to be careful with multi-band metadata to avoid inconsistencies between metadata files. Due to that situation different tools or proceedings have been designed for the application.

The first important utility that must be implemented in this conceptual schema is the possibility to add and delete bands from a multi-band dataset. Connected to this procedure, it is very important to carefully control metadata loses or inconsistencies.

The general value of metadata entries (in that ones it applies, see 2.1 *Multi-band and related topics*) have to be recalculated at the moment a band value changes or when a new band is added or removed.

In the same direction, if a new band is added to a multi-band dataset, its previous metadata file have to be deleted (because that metadata are from now on stored in the multi-band metadata file). Also if a band is emancipated from a multi-band dataset to form and independent data set, its metadata file must be generated according to the metadata previously stored in the multi-band metadata file.

When saving changes in a metadata file, and to avoid repetitions, metadata entry band values are only written if the band value is different from the general value (that it is always written on the file).

Another tool developed to facilitate the management of multi-band data set is the "Apply to all" option. That option allows the user to delete all the specific band values in a metadata entry (all bands values are lost and they change to the same general value). This is a very interesting option in remote sensing images captured by multi or hyper-spectral sensors.

3.2   Geometry implementation

The interdependency between pixel size, bounding quadrangle coordinates and column and row count may be used to calculate some of that unknown values. For example, column count may be deduced from minimum abscissa coordinate, maximum abscissa coordinate and pixel size in abscissa direction.

Depending on the source of the metadata you are documenting or depending on the software implementation users work with, sometimes coordinates are referred to a center of the pixel or to the extern boundary of the pixel. You may use whichever coordinates you know to document the bounding quadrangle just by selecting the appropriate radio-button.

"Apply to all" option must be specially considered in this context because of the special nature of some data file formats. For example in JPEG file format you may not change column and row count (because they are written in the file header) so it has no sense to apply other band values to a JPEG band. This is why "Apply to all" option never affects to JPEG bands.

A caption of the software implementation window to document and calculate those metadata entries is shown in Figure 2.

3.3 *Quality and lineage implementation*

Quality parameters, indicators and measures may be a free text metadata entry but, as said before, software could give some options to be used by different users in the same context. That could be understood as a profile of existent standards.

Our implementation of quality parameters, indicators and measures give some parameter, indicator and measure names that are predefined (or normalized). The user may choose one of them or may use the "free text" option to document any parameter, indicator or measure name he wants.

Due to the tree-list nature of our quality parameters implementation predefined names are not just a list of parameter, indicator and measure names, but are a hierarchic group of names. For example, "Comission Error" should be a normalized measure if it depends on a "Comission" indicator depending on a "Semantic Accuracy" parameter, but it should not if depending on another indicator.

To improve consistency among different software modules, it is necessary that all the related modules considering quality parameters properly document them using this schema and the normalized names. For example, a geometric transformation module will use the "Positional accuracy" parameter, "Horizontal relative accuracy" and "Vertical relative accuracy" indicators and "RMS", "RMS on X",... measure names.

Another important aspect related to quality parameters and multi-source processes is the implications that it may have in other software modules. For example, a mosaicking process should consider which are the quality parameters, indicators and measures related to relative horizontal accuracy of the resulting data set if different sources have different values of positional accuracy measures. As a general statement, always the worse quality measure value existing in source data sets is adopted to be representative of the resulting dataset but other solutions could be considered. Some of the normalized parameter, indicator and measure names can be found in Fig. 4.

```
Parameter: Positional accuracy
    Indicator: Relative horizontal accuracy
        Measures:
            RMS
            RMS on X
            RMS on Y
            Average error for control points,…

Parameter: Completeness
    Indicator: Omission
        Measure: Omission error
    Indicator: Comission
        Measure: Comission error
```

Figure 4: Normalized parameter, indicator and measure names

4   CONCLUSIONS

Due to the complex metadata standards and the need to have useful and simple tools to manage metadata it is necessary (and possible) to find mechanisms to simplify metadata entries but with the same potentiality as the initial standard.

That simplification is much more easy if we consider a special users community with its particular needs and uses. All this simplifications must be only a way to alleviate effort in documenting metadata, not a limitation of the standard potentialities.

ISO international standard is supposed to became the general accepted standard, so a final metadata model and software implementation should consider the final version of that document.

Metadata for Remote Sensing datasets should consider special metadata extensions including metadata entries to document special topics needed in a remote sensing data context.

REFERENCES

Federal Geographic Data Committee. 1998: *Content Standard for Digital Geospatial Metadata.* CSDGM Version 2: FGDC-STD-001-1998. Washington, D.C. Version 2, June 1998.

Comité Européen de Normalisation. 1998. *Geographic Information - Data description – Metadata.* ENV 12657:1998. October 1998.

Federal Geographic Data Committee. 2000. *Content standard for digital geospatial metadata: Extensions for Remote Sensing Metadata (Public Review Draft).* Washington, D.C. Internal Committee Draft, version 0.34, December 21, 2000.

International Organization for Standardization. 2001. *Draft International Standard: Geographic information – Metadata.* ISO/DIS 19115. International Organization for Standardization, Technical Committee 211, September 2001.