EURAC SDI: A Near Real Time and Offline Automatic Metadata Generation Processing Chain

Armin COSTA, Tania P. FIRDAUSY, Markus INNEREBNER, Roberto MONSORNO
EURAC – Institute for Applied Remote Sensing, Bolzano/Italy · support.rs@eurac.edu

This contribution was double-blind reviewed as extended abstract.

Abstract

Scientists dealing with geospatial information usually work with huge sets of heterogeneous geographic data derived from different sources. Without a well-defined and organized structure they face problems in finding and reusing existing spatial data. Due to the increasing amount of collected data, the risk of data redundancy arises, which may cause data inconsistency, space issues and search difficulties. A spatial cataloguing system can facilitate a more efficient spatial data search as well as allowing data exchange with different institutions. Our proposed solution is implementing a spatial cataloguing system along with an automatic rule-based approach metadata generator that processes remote sensing data in Near Real Time (NRT) and simultaneously derives metadata. This paper will further describe how to extract the relevant metadata from the processed data and how we converted this heterogeneous metadata information into a common standardized format. A real-world scenario applied in The European Academy (EURAC) Research Institute for Applied Remote Sensing (IARS) illustrates the procedure of data processing and metadata generation.

1 Introduction

A spatial catalogue is a repository of metadata that provides services to explore, browse and query geospatial data. The main concern is that currently most geospatial data comes with metadata in different formats or sometimes without any metadata, the latter usually applies to GIS data. Spatial metadata can be created and updated manually or with semi-automatic and automatic approaches (OLFAT et al. 2010).

The most popular methods for extracting metadata are hand-coded rule-based parsers, and machine learning (HAN et al. 2003). For highly structured tasks, rule-based methods are easier to implement (OLFAT et al. 2010). Rule-based methods, rule discovery or rule extraction from data, are data mining techniques aimed at understanding data structures, providing comprehensible description instead of only black-box prediction (DUCH 2011). Meanwhile the machine learning methods are approaches that include training data and machine self-correction based on errors in machine performance against the training set (GREENBERG et al. 2006).

The paper presents an automatic approach based on the hand-coded rule-based method that generates metadata in a standardized format extracted from a heterogeneous set of spatial

Jekel, T., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2013): GI_Forum 2013. Creating the GISociety.

data. In addition, since the IARS processes a vast amount of raster data coming from EURAC's satellite receiving station, this paper explains the SDI system architecture which currently generates Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) based products in near real-time together with its standardized metadata.

2 Heterogeneous Data Handling

The data repository of EURAC's IARS consists of data collected from ground and remote sensors and of second-level data produced internally by its own algorithms based on scientific models. The main challenge is how to manage heterogeneous data concerning the improvement of the organization and search as well as the prevention of duplicates.

The EURAC Receiving Station receives on a daily basis raw data from NASA missions Aqua, Terra and Suomi NPP. Beside the NRT scenario, the institute's research deals with many different satellite data: LANDSAT, RapidEye, ENVISAT, and Quickbird to name a few. As the amount of data is rapidly growing, there is the need to automatize data handling and metadata generation.

In order to practically solve the issue of handling data, a concept of data ingestion and data handling has been implemented through the development of our core server, denoted Data Exchange Server (DES), which can be considered as a general, multi-tasking application that performs any type of configurable task or jobs. The tasks and jobs directly involved in the metadata generation will be described in the following sections.

3 System Architecture

The concept and architecture for metadata generation has been developed in order to fit into a previously implemented system for the NRT processing chain at EURAC. The data acquired and processed should be directly integrated into the SDI and immediately become available to local authorities (i.e. civil protection). There are two key requirements which lead to the proposed solution: i) allowing flexibility in generating ad-hoc metadata information for particular purposes; and ii) having interoperability versus third party or custom metadata formats. Further, the concept has been developed considering the necessity to generate on demand (offline) metadata information for existing, third party heterogeneous datasets. The technology used for the implementation of the overall architecture and data processing chain is Java, Perl, XML and XSLT.

3.1 The Near Real Time Data Processing Chain

The NRT data processing chain (figure 1) currently processes MODIS and VIIRS data into user driven application products together with standardized INSPIRE compliant metadata. The core function of the NRT data processing chain lays in the processing nodes that process the data acquired from the receiving station and creates the metadata. A processing node has three essential components which are: i) the Generic Processor (GP); ii) the Wrapper; and iii) the Processor. All the three components get the instructions through a centralized XML configuration file. The GP's main function is to perform the pre- and post-processing of a product. The pre-processing consists of setting up the environment, loading the

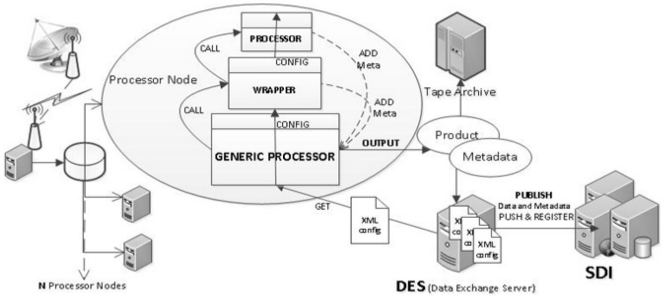

Fig. 1: EURAC NRT Processing Chain

XML configuration file, checking requirements and parameters, and basically controlling the whole data processing chain. The post-processing includes metadata and quick look image generation. The wrapper's main functionality is to call and activate the processor and to further abstract the processing ensuring flexibility. Finally the processor processes the data according to a specific algorithm. The metadata produced by the GP is transformed into a standardized form, namely INSPIRE, by the DES. Besides transforming the metadata, the DES also distributes the data and metadata into the SDI.

3.2 The Automatic Metadata Generation Method

The method implemented within the proposed architecture (figure 2) considers two scenarios for the metadata generation: the NRT and the Offline (on demand) generation. Both scenarios are based on the hand-coded rule-based parser method, since the intermediate metadata file originates either from extracting the required metadata elements embedded in the data (HDF) files or by parsing them from the processor configuration file. The parsing is based on specific hand-coded rules which are implemented as plug-ins. Furthermore, the intermediate metadata parsing and final metadata format generation is based on specific hand-coded rules in form of XSLT transformations which are implemented as dedicated runnable plug-ins (for each metadata type) on the DES server application.

3.2.1 NRT Generation

The processing architecture provides an abstraction layer stack defined as GP and each processor is further abstracted by means of a Wrapper interface. The concept of the method is that metadata can be added statically and dynamically at each level of the abstraction layer stack. This provides a high level of flexibility allowing to add custom metadata when required, which enables us to follow particular specifications like the INSPIRE metadata directive. Once the processing of a given product within the processing chain is completed, the GP gains control again and generates the final metadata, which consists of an "Intermediate Metadata Format" XML file and an image preview and thumbnail. Next, the GP also forwards the metadata and data to the DES server. The DES server can be configured to further process the "intermediate metadata format" and generate one or n-dedicated formats as required for example by the catalogue service. The DES also takes over the task to feed and register the data and metadata into the SDI (i.e. Catalogue and Map Service).

3.2.2 Offline Generation

This method provides a way to generate metadata for different data types (i.e. Landsat, RapidEye, CSK etc.). In this scenario the metadata generation is performed on demand in a subset of the NRT architecture, directly via DES server. The DES provides a way to load and execute dedicated plug-ins in form of task implemented within the framework. A dedicated plug-in generates directly the "Intermediate Metadata Format", the rest of the process is than performed in the same way as for the NRT generation method, transforming the metadata to some destination formats and transfer them to the destination.

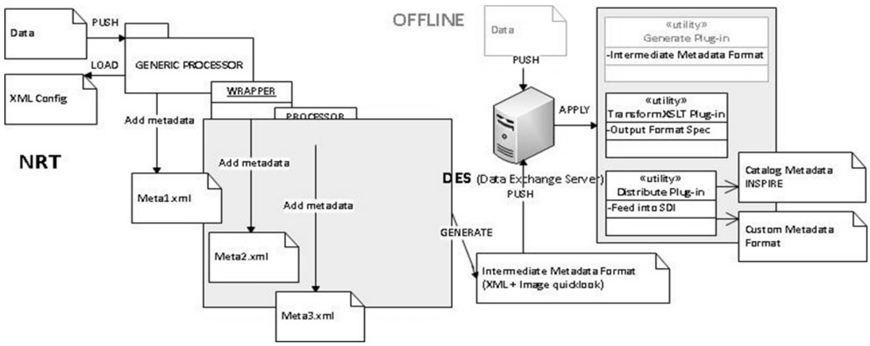

Fig. 2: Metadata Generation Architecture

The choice to have an "Intermediate Metadata Format" is based on the requirement to be able to provide different kinds of metadata formats, and by the necessity to re-generate the metadata with added information.

4 Conclusion and Outlook

Metadata is one of the most important information associated to spatial data because it provides vital information about data identification. We implemented an automatic method based on the rule-based approach that extracts metadata from a heterogeneous set of sensor data. The automatic metadata creation method was developed essentially to support the NRT data processing chain which produces MODIS and NPP based products. In addition to that we implemented a process that generates metadata for offline data. Recently these functionalities have been deployed as part of our SDI in our research institute with success. Future work includes the development of dedicated plugins to support the automatic metadata generation for other types of satellite data.

References

DUCH, W (2013), Rule-Based Methods. Encyclopedia of Systems Biology, DUBITZKY, E. et al. (Eds.). Springer 2013 (in print, accepted 2011).

GREENBERG, J., SPURGIN, K. & CRYSTAL, A. (2005), Final report for the AMEGA (Automatic Metadata Generation Applications) project. http://www.loc.gov/catdir/ bibcontrol/lc_amega_final_report.pdf.

HAN, H., GILES, C. L., MANAVOGLU, E., ZHA, H., ZHANG, Z. & FOX, E. A. (2003), Automatic Document Metadata Extraction using Support Vector Machines. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, 37-48.

OLFAT, H., RAJABIFARD, A. & KALANTARI, M. (2010), Automatic Spatial Metadata Update: a New Approach. Paper presented at the FIG 2010, Sydney, Australia.