Earth Sci Inform (2017) 10:55–68
DOI 10.1007/s12145-016-0276-0

RESEARCH ARTICLE

Organizing earth observation data inside a spatial data infrastructure

Markus Innerebner[1] · Armin Costa[1] · Ekaterina Chuprikova[1] · Roberto Monsorno[1] · Bartolomeo Ventura[1]

Received: 6 July 2015 / Accepted: 18 October 2016 / Published online: 9 November 2016

Abstract Scientists as well public institutions dealing with geospatial data often work with a large amount of heterogeneous data deriving from different sources. Without a well-defined, organized structure they face problems in finding and reusing existing data, and as consequence this may cause data inconsistency and storage problems. A catalog system based on the metadata of spatial data facilitates the management of large amount of data and offers service to retrieve, discover and exchange geographic data in an quick and easy fashion. Currently, most online catalogs are more focusing on the geographic data and there has been only few interests in catalogizing Earth observation data, in which in addition the acquisition information matters. This article presents an automatic metadata extraction approach that creates from different optical data deriving from various satellite missions of scientific interest (i.e. MODIS, LANDSAT, RapidEye, Suomi-NPP VIIRS, Sentinel-1A, Sentinel-2A) metadata information, based on an extended model of the standard ISO 19115. The XML schema ISO 19139-2 with the support of gridded and imagery information defined in ISO 19115-2 was examined, and based on the requirements of experts working in the research field of Earth observation the schema was extended. The XML schema ISO 19139-2 and its extension has been deployed as a new schema plugin in the spatial catalog Geonetwork Open Source in order to store all relevant metadata information about satellite data and the appropriate acquisition and processing information in an online catalog. A real-world scenario that is productively used in the EURAC research group institute for Applied Remote Sensing illustrates a workflow management for Earth observation data including data processing, metadata extraction, generation and distribution.

Keywords Spatial data infrastructure · Spatial catalog · Metadata extraction · Managing Earth observation data

Communicated by: H. A. Babaie

✉ Roberto Monsorno
roberto.monsorno@eurac.edu

[1] Institute for Applied Remote Sensing, EURAC Research, Bolzano, Italy

Introduction

In the last half decade the use of spatial information has grown in importance mainly in public sectors as well as in different research fields, such as Earth Observation (EO), environmental monitoring and computer science. This area is one of the fastest growing today (Litwin and Rossa 2011) and the global revenue in this sector is 10 % and is growing every year (Litwin and Rossa 2011). The popularization of geographical content on the world wide web has a significant impact, mainly driven by the online services provided by Internet pioneers such as Google, Yahoo, and Microsoft, in which the geolocation takes an important role. Also in the Remote Sensing research field spatial data has a significant importance. Research institutes, such as NASA or DLR, that daily receive from the receiving station a huge amount of satellite data in heterogeneous formats, need to manage this information efficiently, in order to avoid having space problems, inconsistency because of data redundancy and other kind of problems provoked by the quantity and diversity of data.

The very first objective of the institute for Applied Remote Sensing is to focus on the integrated environmental monitoring in mountain regions. Therefore the scientists are dealing with different kind of vector and raster data acquired from different sources (satellite, ground sensor, web, partners, third-party providers). This information is stored in different repositories (file-servers, databases, data-tapes) and the big challenge consists in organizing this large amount of heterogeneous data in an efficient, reliable, robust and intuitive way.

A Spatial Data Infrastructure (SDI) that provides services for storing, discovering, querying and exchanging geographic information, simplifies the management for large and heterogeneous data. For retrieving information, spatial data can be managed in a similar way as books are organized in a library, using a catalog. A spatial catalog is a system that allows to search for spatial data based on metadata and it is a component of the SDI. The term metadata refers to *"data about data"*. At the basic level, metadata should characterize an entity by answering at least the following questions (Litwin and Rossa 2011): *What* — what it is and what it refers to; *Why* — to what purpose it was created; *When* — when it was produced, published or updated; *Who* — who created or developed it; *How* — how it was produced, is it reliable, how to get access on it; *Where* — what area it refers to. Wilson (2008) defines three different levels in the usage of metadata: (a) *discovery* metadata: discover the semantics of a data element in data sets (metadata scanning); (b) *exploration* metadata: comprises detailed information about quality, accuracy and origin of the data; (c) *exploitation* metadata: expresses how to read, transfer and integrate this data in applications.

The benefit in using metadata is a quick access to the desired information, analog like the book search in a library using a catalog. A spatial catalog offers to search for any kind of spatial data independent on the location where it is stored. Without a catalog it would not be possible to perform a search on two datasets coming from different sources (e.g. filestore vs. database). For a spatial catalog metadata provide information about the purpose, quality, actuality and accuracy of a spatial entity set and consequentially provide the relevant input parameter for an accurate search. Moreover a catalog allows to granularize the search including a keyword based search, temporal and spatial filters and to manage user policies.

Most catalogs use a standardized format for structuring their metadata. The standard technical committee established within ISO (ISO/TC 211) in cooperation with other organizations developed the series of International Standards and Technical Specifications in the range starting at 19101 (Ostensen and Danko 2005). For the organization of spatial metadata the following ISO standards have a significant importance: ISO 19115 provides an abstract and logical model for geospatial metadata. Because of its international acceptance, this standard has become a part of the "OpenGIS Specification" as the abstract model for the management of metadata. ISO 19115 defines the metadata to record information about the identification of geospatial data sets, possible approaches for distributing the data, details about the quality, geographic facets about the coordinate reference system and the coverage, temporal information about the date of acquisition, the owner of the data.

However, this standard can not fulfill a big part of the requirements for imagery and gridded data that are relevant for EO-data. In order to cover these missing information in a standardized format, there was published the extension for gridded and imagery data (ISO 19115-2) in the year 2009. That new standard details, how a dataset was acquired, what kind of instrument has been used to produce the data, what kind of algorithm was used to process the data and remarks about the quality of a dataset.

Since XML became as one of the most frequently used exchange format, there has been developed a corresponding XML encoding for these two abstract models. The standards ISO 19139 being the implementation of ISO 19115 and ISO 19139-2 being the one of ISO 19115-2 consist of a set of XML schema files that define the grammar of the structure.

There exist some online catalog products, such as the ESRI Geoportal Server or Geonetwork Opensource, that use some of these ISO standards for managing metadata. Both products are Open Source and offer predefined metadata schema templates fulfilling ISO 19139. But none of these products support to organize metadata of ISO 19139-2.

This paper describes how large and heterogeneous amount of EO-data coming from different sources can be efficiently organized in a centralized SDI using a spatial catalog. Due to the vast data volume an automatic data deployment approach is necessary to register datasets in the SDI, to extract metadata from the raw data, to populate it in the catalog and to archive data in a store.

The contribution of the actual article includes:

- an implementation of an automatic metadata extraction approach based on methodology of the hand-coded rule-based parser that generates metadata fulfilling the ISO standards,
- an extension of the metadata standard ISO 19115-2 and ISO 19139-2 to cover specific details about imagery and gridded data,
- a customization of an Open Source spatial catalog that allows to organize EO-data inside it.

Nowadays, with this enhancement it is possible to manage all the institute's relevant satellite products within the SDI.

The paper is organized as following. Section Related work delineates related work about SDI implementations, search catalogs and metadata generation methods. Section Processing of earth observation data and metadata generation describes the architecture for managing EO-data including near-real-time processing, the metadata extraction and creation process. In Section Extending and implementing ISO schema for EO data are presented the developed metadata models suitable for EO-data and how its integration in the spatial catalog. Section Conclusion and outlook summaries the paper and points out ongoing and future activities.

Related work

A geoportal is a type of web portal used to discover, view and access spatial information using geographic services (display, editing, analysis, etc.) via the Internet. Geoportals are important for an effective use of geographic information systems (GIS) and a key element of Spatial Data Infrastructure (SDI). Geographic information providers, including government agencies and commercial sources, use geoportals to publish geospatial metadata. Geographic information consumers, professional or casual, use geoportals to search and access the information they need. As a consequence geoportals serve an increasingly important role in the sharing of geographic information, and they can assist in avoiding duplicated efforts, inconsistencies, delays, confusion, and wasted resources.

The most common online geoportals are the ESRI Geoportal Server and Geonetwork Opensource. Both of them are Open Source products, offering predefined metadata schema templates for ISO-19115 and ISO-19139, as well as the possibility to define and create customized extensions. Furthermore, Geonetwork offers a large amount of documentation, examples and templates. Geonetwork offers a predefined set of schema plugins based on ISO standards for describing different kind of data. The standard technical committee established within ISO (ISO/TC 211) in cooperation with other organizations developed the series of International Standards and Technical Specifications in the range starting at 19101 (Ostensen and Danko 2005).

Previously there have been elaborated general metadata standards. In particular for the geographic area the Spatial Data Transfer Standards (SDTS), the Vector Product Format (VPF) and the Digital Exchange Standards (DIGEST) were developed to allow the encoding of digital spatial data sets to be exchanged in spatial data software. These standards include the support of metadata, but have not been considered until recently to standardize their encoding for the export of the data. On the contrary, ISO 19115 defines metadata elements and schema in order to document geographical data in standardized and comprehensive way. The XML Schema implementation derived from ISO 19115 could be encoded according to ISO 19139.

Considering the fact that the institute mainly works with the Earth Observation data, the ISO 19115 could not fulfill all the requirements for metadata description. For this reason the extension ISO 19115-2 and its XML schema implementation ISO 19139 were taken into account. Thus ISO 19115-2 extends the existing geographic metadata standard by defining the schema required for describing imagery and gridded data (National Coastal Data Development Center et al. 2012b). According to the this extension, the acquisition information could be described including platform and instrument characteristics, information about environmental condition, plan of the measurements and other details about the acquisition process. Despite the standard covers a great deal of the information, the diversity of geographic data causes difficulties in satisfying all the requirements. Hence the ISO metadata standard provides a standardized way for users to extend their metadata and still ensure interoperability allowing other users to comprehend and exploit this extended metadata (Ostensen and Danko 2005). Under the analysis of the Earth Observation data it was revealed that imagery information obtained from different satellites could not be fully defined within the framework of ISO 19115 and its extension ISO 19115-2. Thus it was decided to work on extension for ISO 19115-2 which can fully cover any scientific interesting satellite mission, optical and SAR, such as Landsat, MODIS, S-NPP, RapidEye, Sentinel-1 and Sentinel-2.

Currently, the main concern is that most geospatial data comes with metadata in different formats or sometimes even without metadata, the latter usually happens to vector data. Spatial metadata can be created and updated manually or with semi-automatic and automatic approaches (Olfat et al. 2010). The first two approaches are considered as monotonous, time consuming, and a labor-intensive processes by organizations and they are commonly viewed as an overhead and extra cost (Olfat et al. 2010). Much research has been focused in the field of automatic metadata extraction in particular in the research field of digital libraries where interoperability is essential for exchanging articles between different institutions (e.g. IEEE). There have been used several methods for automatic metadata extraction: regular expressions, machine learning, and rule-based parsers are the most popular (Giles et al. 2003). Regular expressions techniques filters out metadata from the data using specific patterns. 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. 2005). Machine learning methods are robust and adaptable, and theoretically can

be used in any document set (Olfat et al. 2010). For information extraction they include symbolic learning, inductive logic programming, grammar induction, Support Vector Machines, Hidden Markov models, and statistical methods. Rule-based methods, rule discovery or rule extraction from data are the data mining techniques aimed at understanding data structures that providing comprehensible description instead of only black-box prediction (Dubitzky et al. 2011). In metadata extraction or creation this means a set of rules based on a particular standard are predefined which governs the metadata parsing and creation process. For highly structured tasks rule-based methods are easier to implement (Olfat et al. 2010). Rule-based parsers are straightforward to implement and they depend on a specific application domain and need experts to set up the rules of generation, but compared to machine learning methods they are easier and faster to be implemented. Having in-house experts and due to lack of time the rule-based parser approach has been chosen as metadata extraction method. Basically this method uses XML technologies (XML, XSLT, XPath) for the extraction and generation of geospatial metadata.

Processing of earth observation data and metadata generation

Considering the different kind of data as well as the large amount of daily received data (receiving station, ground sensor, third-party satellite data providers) it was necessary to provide a solution for the issue of automatic handling of data. A concept of data ingestion and data management in general has been implemented through a near real time processing chain as illustrated in Fig. 1.

Following the ideal workflow of one specific dataset, the Data Feeder collects new data, triggers the processing, extracts and generates the metadata. Then it transforms the generated metadata to a standardized format (e.g. ISO 19139-2), stores data and their metadata into the SDI. There are some key requirements which lead to the proposed solution, namely allowing flexibility in generating ad-hoc metadata information for particular purposes, and possibly have interoperability versus third party or custom metadata formats. Further, the concept has been developed considering also the necessity to generate on demand (offline) metadata information for existing, third party heterogeneous datasets. In order to meet these requirements we introduced the concept of an "Intermediate Metadata Format" which consists of a dynamically generated XML file containing all information which can be extracted directly from the data (e.g. pixel reference information, dimensions, etc.) and a statically coded XML file that holds any other needed metadata fields (e.g. INSPIRE metadata). This intermediate metadata format is transformed to different target formats. The technologies used for the implementation of the overall architecture and data processing chain are Java, XML and XSLT.

An important component in the processing chain is the Data Exchange Server (DES). The DES is an application that enables to transfer data among different systems. It was developed to fit the requirements of the heterogeneous data handling, like ingestion, processing and transferring from one system to another. The DES also allows to deliver data to external users by implementing a concurrent, multi-threaded mechanism to transfer different kind of data in PUSH and GET mode using standard protocols such as SFTP, FTP, SSH, WEBDAV. In general the DES might be considered as a multi-tasking application that can perform any kind of configurable tasks or jobs by executing dedicated plugins.

The task and jobs directly involved into the metadata generation will be described in the following sections. Section The near real time metadata generation will focus

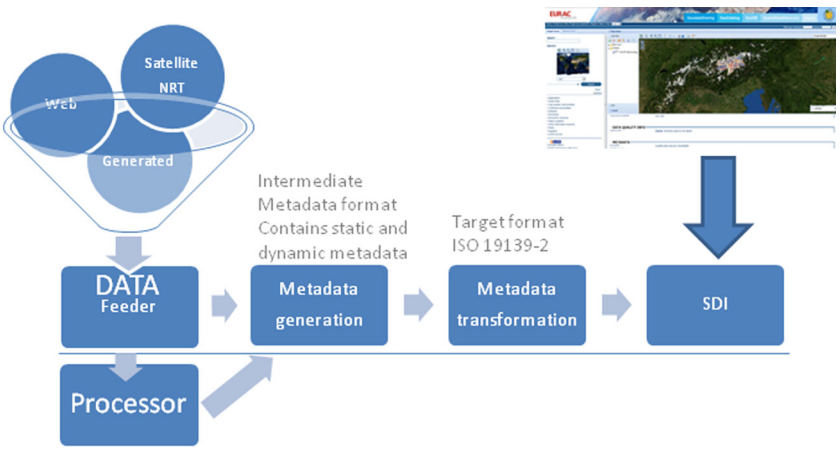

Fig. 1 EURAC Processing Chain

on the NRT processing chain metadata generation used for data acquired directly by the receiving station or downloaded from the ESA Sci-Hub. Section Offline metadata generation describes the offline metadata generation used for third-party data.

The near real time metadata generation

The core of the NRT data processing chain consists of several compute nodes that implement the Eurac Generic Processing Framework (EGPF) as depicted in Fig. 2. It has three essential components, namely the Generic Processor, Wrapper and Processor.

The main functionality of the Generic Processor is the pre- and post-processing of the products. The pre-processing includes: (a) setting up the processing environment, (b) loading the XML configuration file, (c) deploying on-the-fly the processor application, (d) checking the processing requirements and parameters, (e) extracting the available metadata from the processor input files, and (f) invoking the selected processor by calling the Wrapper interface. The post processing includes the generation of the metadata and the creation of a preview image of the final product. A configuration file provides the relevant information needed for processing the data. This includes the used algorithm with its parameters, auxiliary files and static metadata properties.

The Wrapper component was introduced to call and activate the specified processor and to further abstract the processing chain which ensures a higher flexibility and a plug-and-play nature of the algorithm.

The Processor component implements the specified algorithm and produces a higher level product out of raw satellite data together with some preliminary metadata that are delivered to the Generic Processor.

The core concept behind the metadata generation method adopted, is the possibility to add statically and dynamically metadata at each level in the EGPF abstraction layer stack. This provides a high level of flexibility for example if the output of a given processor shall include also more detailed, customized metadata information required for example for a research or further processing purposes. This flexibility provided allows also to be able to follow some particular specifications like the INSPIRE directive for metadata.

As example a product that generates an enhanced snow coverage map using MODIS data including a "CLOUD coverage indicator" could be considered. In this case, the processor itself can include this metadata property via a dedicated interface without involving the Wrapper and Generic Processor layers.

Once these data and metadata are passed back to the Generic Processor, final metadata, consisting of an "Intermediate Metadata Format" XML file, an Image preview and thumbnail, are generated.

Furthermore, the Generic Processor forwards the metadata and data to the DES. The DES can be configured to further process the "Intermediate Metadata Format" and generate one or more different dedicated target metadata 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) as described in Fig. 3.

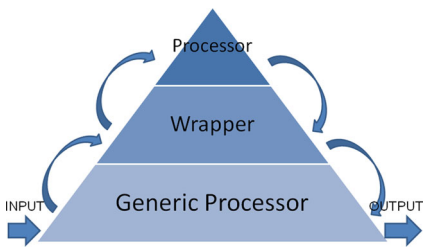

Fig. 2 EGPF

Offline metadata generation

To facilitate the insertion of metadata from third-party data, there has been introduced a simple and automatic workflow to store data and related metadata in the EURAC SDI. In this scenario, the metadata generation is performed on demand in a subset of the NRT architecture, directly via DES instead of via the EGPF. The DES allows to load and execute a dedicated plugin in form of a task. Therefore the Offline metadata generation is a plugin of the DES that generates directly from the input data the "Intermediate Metadata Format" without passing through the EGPF. The rest of the generation process is identical to the NRT.

During the integration of a new dataset, all available metadata information are extracted, amplified with mandatory ancillary information (e.g. URI) and stored in an intermediate file. Afterwards, this file is merged with the static metadata file, provided by the DES, and transformed to the target format using XSLT transformation. The output of this process represents the metadata end-product implementing the geospatial metadata ISO standards, and it is inserted into the spatial catalog.

A possible data insertion workflow looks like the following. A user obtains a new dataset from a third-party data provider (i.e. NASA or ESA). In order to register this dataset in the SDI, it is necessary to put the data into the appropriate data feeder. Once the data is there, the DES application periodically checks for new data and

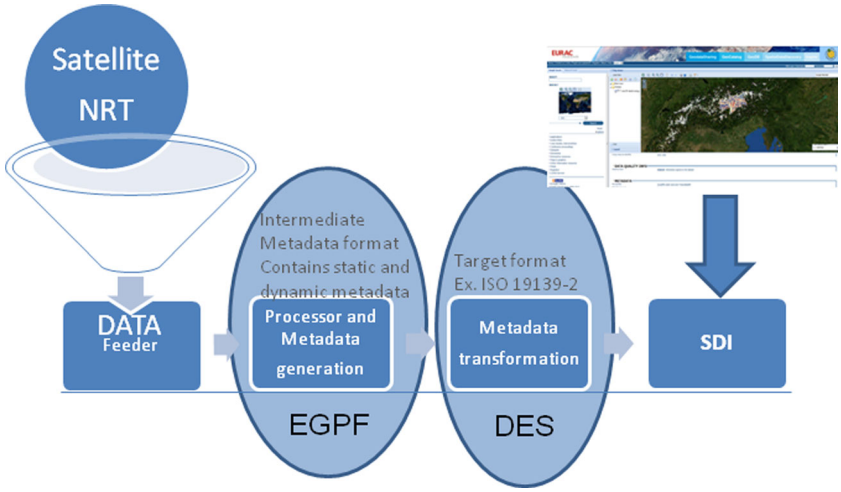

Fig. 3 NRT Metadata Generation

starts the process of metadata extraction, integration and transformation. Figure 4 shows an excerpt of the metadata generation. From the IMF the acquisition date that is related to the data is selected via an XSLT expression (XPath). Correspondingly, the metadata category field is selected from the static metadata file. The result of the XSLT transformation is the target format. Once the target format is generated, the DES will register data and its metadata in the SDI.

The choice to have both, an "Intermediate Metadata Format" and a static metadata file, allows a lightweight and fast regeneration of the target metadata file without reloading the entire input data. Further this approach will not require any particular programming skills, when any metadata definition changes. This means, by simply having a knowledge in XML technologies (XML schema and XSLT) a non programmer is capable to produce new metadata output.

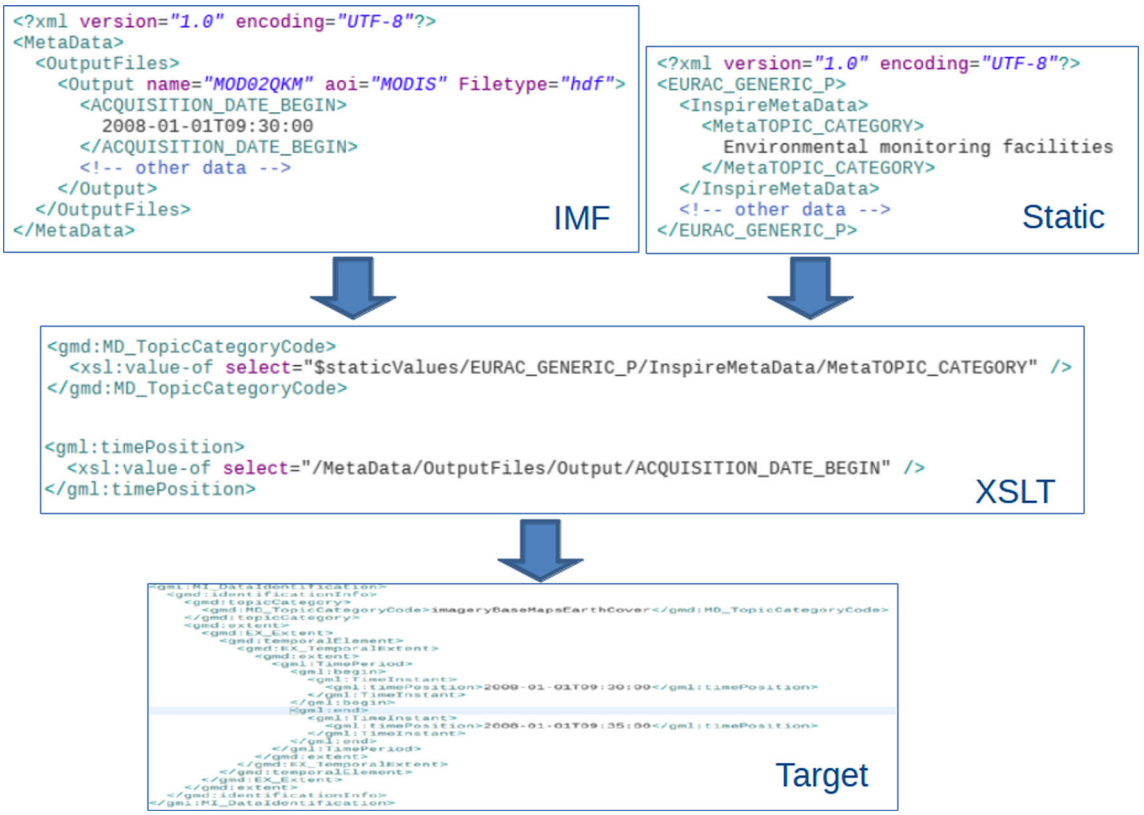

Fig. 4 Metadata Transformation

Extending and implementing ISO schema for EO data

As already mentioned in the introduction Section Introduction there exists the standard XML encoding ISO 19139 for GIS metadata and its extension ISO 19139-2 for imagery and spatial data used in particular for EO metadata. Nevertheless these metadata standards can not fulfill all requirements of EO scientists.

For instance, for a refined search the following queries could not be answered:

- *Find a specific dataset that was acquired from satellites flying in ascending orbit with a cloud coverage with less than 10 %.*
- *Find Landsat images with a spatial filter bases on the worldwide reference system (WRS) path within a particular WRS path and row that can fulfill the conditions of the mask availability.*
- *Find all RapidEye data captured with a specific license type or to find all RapidEye data with particular information about visibility, aerosol type or water vapor..*
- *Find all Sentinel-2A data within a certain period, passing over a certain area with a cloud coverage less than 15 % .*

Moreover any optical sensor based on satellite mission, like for example MODIS, S-NPP VIIRS LANDSAT, RapidEye, Sentinel-2A, have specific characteristics about the single bands of the sensor, that are not mentioned in the existing schema, but they are relevant for researchers. Thus it leads to the necessity of extending the metadata standard 19139-2 with new classes. The following table gives a review about the existing standards, how they are implemented as an XML encoding and how they are supported in the Open Source catalog Geonetwork (Table 1).

The metadata standard for geographic data in the abstract model ISO 19115 has the XML encoding ISO 19139 that is fully implemented in XML and completely integrated as a schema plugin in Geonetwork (tested with version >2.6). Its extension ISO 19115-2, the metadata standard for imagery and gridded data, with the XML encoding ISO 19139-2 is not fully implemented in XML (missing packages and inconsistency according to the the abstract model) and there is not an available schema plugin in Geonetwork. For this reason EO-data can only be stored in the catalog with the basic information from the abstract model ISO 19115. Naturally, for the new introduced extension ISO-19115-2e until now there is not an available abstract model, XML encoding and a schema plugin for Geonetwork. How these missing gaps are filled is described in the next two sections.

The extension of ISO 19139-2

Basically the EURAC enhancement extends the packages MI AcquisitionInformation, MI ContentInformation, MI SpatialRepresentationInformation and MI ReferenceSystemInfo. For a clear separation the new introduced classes start with the prefix *MIE* that is the short name for Metadata Imagery EURAC.

For a comprehensive depiction of the dataset the package MI AcquisitionInformation was extended with the following classes and properties as shown in Fig. 5.

The class MI Platform that is the physical satellite platform, was enhanced with the platform name, that identifies the name of the satellite platform and the orbit type, which is given according to altitude classifications for geocentric orbits, for instance Low Earth Orbit. The class MIE Instrument derived from MI Instrument was enhanced with two properties that describe the type of scanning system used by the sensor and the spatial resolution of the sensor. The scanning system defines the way of how the data was captured along or across the track. Meanwhile the resolution provides the information of the size of the ground objects distinguished by sensors. Thus the same area can be captured with different resolution. The choice of the resolution depends on the researcher's goal.

The class MI Operation describes the relevant information during the operation phase of the acquisition, was enlarged with information about some measurement characteristics as spacecraft view angle, orbit direction, and equator crossing time. Since the researchers of the institute asked to query if an image was captured during day or night the property dayNightFlag was introduced. Regarding equator crossing time (ECT), the value of this element remains nearly constant throughout the year, however the ECT data may be used for global vegetation monitoring. The Spacecraft view angle was introduced to support the researchers with information concerning viewing angles.

Table 1 Support of ISO standards

ISO (abstract model)	ISO (XML encoding)	Extension of	XML available	Schema plugin in GN	Prefix
19115	19139	—	yes	yes	MD
19115-2	19139-2	19139	incomplete	no	MI
19115-2e	19139-2e	19139-2	no	no	MIE

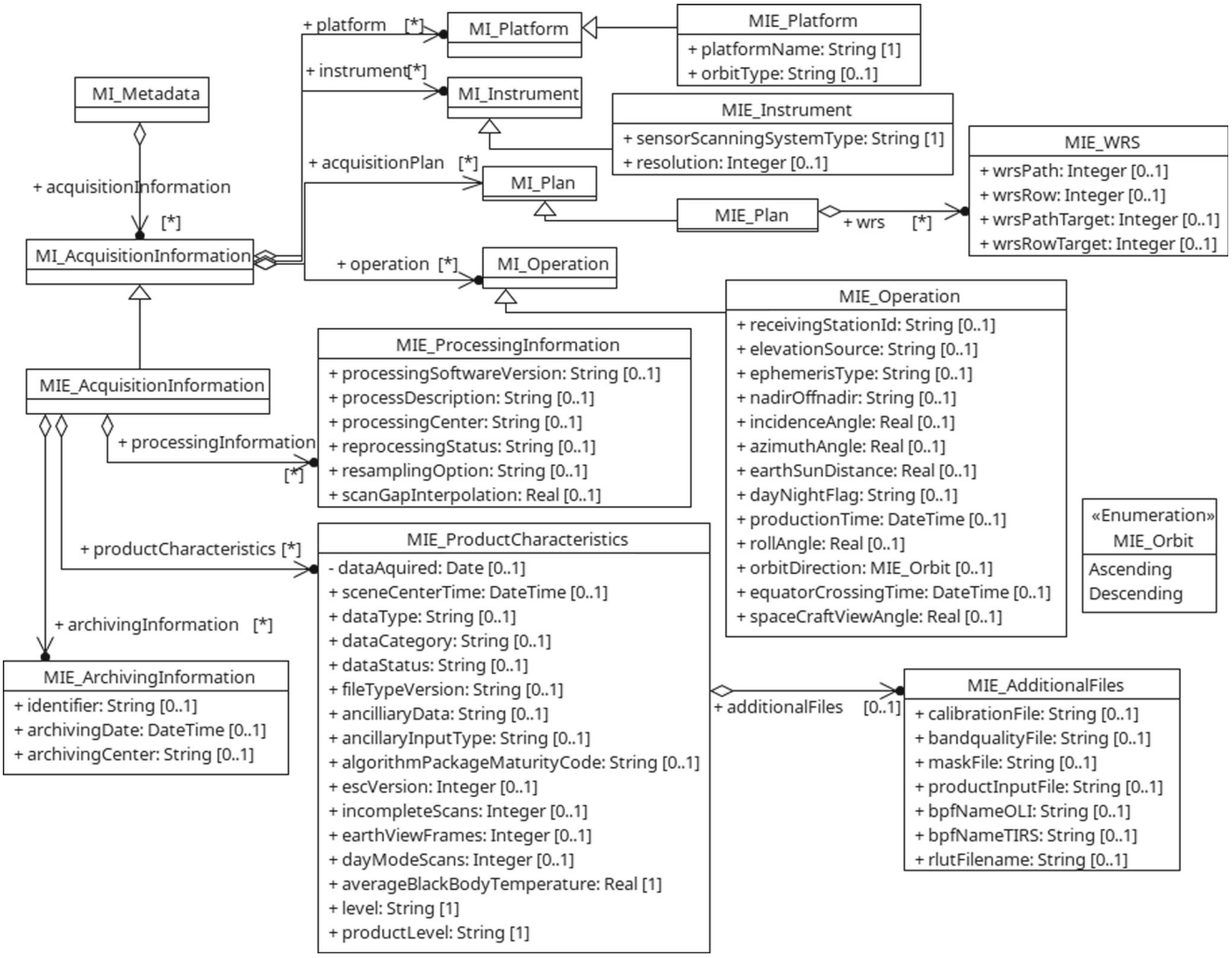

Fig. 5 The *acquisition information* package in the extension of ISO 19139-2 portrays more details of the operation processes during the acquisition phase and auxiliary information about the acquisition process of multi-spectral sensor data

So that immediately it can be seen that large viewing angles provide enhanced sensitivity to atmospheric aerosol effects and to cloud reflectance effects, whereas more modest angles are required for land surface viewing. Besides the element Nadir/Off Nadir provides with the information if sensors looked vertically downward (at nadir) or sideways. The element Elevation Source supplies with the information about the digital elevation dataset that was used for terrain correction of the product. Apart from the mentioned elements the class was extended with following elements: receiving station, ephemeris type, incident angle, azimuth angle, earth sun distance, production time, and roll angle.

Moreover it was necessary to introduce some new classes, in order to reflect the metadata which could not be placed in the framework of the existing classes. The class MIE ProcessingInformation was introduced to define specific information about image processing including processing software version, process description, processing center and reprocessing status. Besides it includes the information about resampling option that defines the method used to produce the image. Moreover, the class MIE ProductCharacteristics was created and it includes the auxiliary information about data type, data category, data status, file type version and additional files. As an internal important search criteria the new properties level and product level were introduced to classify EO-data in different levels (LEVEL-0, LEVEL-1, etc.) and product levels (MOD 01, MOD 02). The property Averaged Black Body Temperature usually takes place in the metadata attributes of such sensors such the NOAA-AVHRR, ERS-ATSR and TERRA-MODIS, that are equipped with a band detecting infrared

radiation. Besides, based on the class MIE AdditionalFiles the researchers can find the references to supplementary documents, describing calibration, mask, band quality, and product input. The class MIE ArchivingInformation was inserted to define archiving date, archiving center and identifier. Hereby the location store of the image is annotated with a timestamp.

Furthermore the package MD ContentInformation was extended, since it was necessary to include additional characteristics for satellite images and for each band. Figure 6 shows the new introduced classes with its properties.

Therefore, the class MI Band was extended to MIE Band with minimum and maximum values for radiance and reflectance, information about masking, values, shifting, binning and corrections applied for every band dimension. Each band can be described according to its type that includes thermal, emissive, panchromatic or reflective bands. The class MI ImageDescription was enhanced to MIE ImageDescription including details about the cloud coverage that reflects a cloud cover assessment confidence and indicator of how the cloud cover percentage has been estimated. Moreover, it includes the percentage of usable data, the pixel format and information about the nodata (missing) values in the raster cells without content. In addition the class MIE ImageGeometry is introduced for detailing the geometry of the image: these properties specify the size of the image, the number of columns and rows that are essential for the navigation through the data set and other particular information for the geolocation of raster data. Moreover some new classes specifying the correction parameters, mask and calibration information were added. The class MIE CorrectionParameters contains details about how the captured image was corrected, MIE MaskInformation can answer the questions about mask availability,

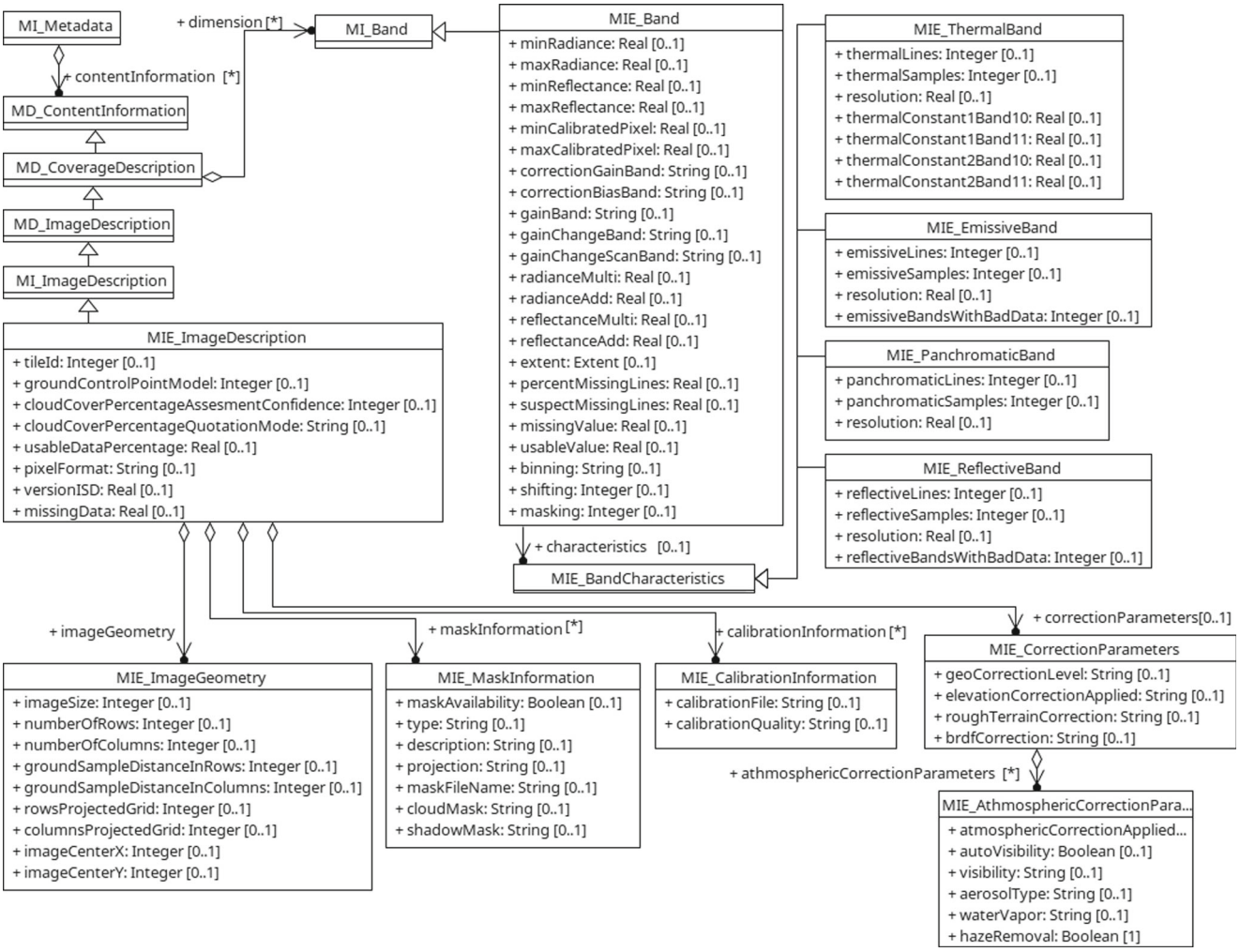

Fig. 6 The *content information* package in the extension of ISO 19139-2 was enhanced with specific characteristics about the individual bands and with auxiliary information about the acquired image

type, description, projection and mask file name and MIE CalibrationInformation includes details how the image was calibrated.

In the package MD SpatialRepresentation the class MI Georectified that is responsible in the transformation process of the image rectification was extended with details about the used orientation during the creation of the image.

Also the information about a specific reference system for Landsat data, denoted as Worldwide Reference System (WRS), was added in the package reference system (MD ReferenceSystem → MIE WRS). This modification allows to search for the Landsat path and row to get the satellite track and row.

Finally in the package Identification Information the aggregation class was enhanced with new properties for describing time series data.

Implementing an XML schema for EO-data

Based on the previously introduced models the implementation of the corresponding XML schema was straightforward. It must be said that until now only few attention has been paid in implementing an XML encoding for EO-data. The XML encoding for ISO 19139 is fully published at the ISO/TC 211 web page (http://www.isotc211.org/2005/gmd), while the extension ISO 19139-2 is only partially published (http://www.isotc211.org/2005/gmi). In the implementation procedure there has been first added missing or incorrect information in the XML encoding of ISO 19139-2 and second introduced a new XML schema for the MIE packages.

During the investigation of the existing ISO 19139-2 schema there has been identified some inconsistency between some properties in the UML schema and the properties in the XML schema. Also the complete package of MI SpatialRepresentationInformation was missed in the XML part. These consistencies were corrected and published on the SDI web server of the institute http://sdi.eurac.edu/metadata/iso19139-2/schema/gmi. This URL can be used for the validation of EO metadata XML files against the grammar defined for ISO 19139-2.

The new packages used for describing metadata of imagery and gridded information with the customized extensions has been added as in a separate structure. The XML schema was published at the URL http://sdi.eurac.edu/metadata/iso19139-2/schema/gmie and it is used as grammar for the validation of most of the metadata files of EO-data used at the institute. As implementing instrument "Eclipse Web Tool Platform" was used that facilitates the development with some important features, such as code completion, integrated navigation to XML schema files, instant validation of the developed schema and the corresponding metadata XML files.

Creating a schema plugin for EO-data in the spatial catalog

A spatial catalog is one of the fundamental components in an SDI, because it provides services for the discovery, browsing, and querying data. Since many projects tenders require to use Open Source technologies – this is also the philosophy of the institute – it was opted for the catalog GeoNetwork Opensource. Furthermore, Geonetwork offers a large amount of documentation, examples and templates. Besides, Geonetwork provides a predefined set of schema plugins based on ISO standards for describing different kind of data. This product is widely spread and successfully used in other different GIS areas for more than five years. Geonetwork Opensource is based on standard technologies (ISO, OGC) equipped with the characteristic to offer extendability in the creation of metadata. The used relational database, characterized by a fixed relational structure, does not affect the flexibility of the metadata. In fact, the use of an XML content, stored as a single text column, allows to store customized elements.

Within Geonetwork Opensource a metadata customization implies to provide a schema plugin. A schema plugin is a directory with XSLT stylesheets, XML schema descriptions and other information that needs the software to index, view and edit metadata. By default the product is installed with a predefined set of plugins including the metadata standard ISO-19139, but without the extension ISO-19139-2. Consequentially there has been implemented two new schema plugins following the guidelines from the Geonetwork developer manual (Various 2013). One plugin comprises all metadata information from ISO 19139-2 and in general can be used for modeling EO-data. The second plugin was developed for internal usage to fulfill the requirements of the scientists inside the institute.

The two new schema plugins are denoted and ISO 19139-2e. One of the critical steps during that phase was the adaption of the presentation view according to the requests and proposals of the researchers. Another crucial part was the indexing via the Lucene Search framework of the relevant search fields that were introduced.

Figure 7 shows two excerpt of the spatial catalog used by the institute. The Fig. 7a shows the results from a search of all satellites images captured during the day. Besides, on the left side a search view offers to performs a simple search or an advanced refined search including spatial and temporal filters. The central right part in the same figure matches the result view including the most relevant information of the found record in combination with a preview. It is possible to inspect the details of such a record and eventually to download the corresponding data. In addition – as illustrated on the right side Fig. 7b – it allows to view the found

dataset inside a interactive geographic map in combination with different base layers (Google Maps, Microsoft Bing, Openstreetmaps). Consequentially, there is no additional GIS software required for viewing such a dataset.

Conclusion and outlook

Metadata is one of the most important information associated with spatial data, because it provides vital information about the identification of a dataset. In this paper there was presented an automatic method based on the rule-based methodology that extracts metadata from a heterogeneous set of remote sensor data. The used metadata formats are based on ISO standards following the XML encoding from ISO 19139 and ISO 19139-2. The metadata standard ISO 19139-2 for describing imagery and gridded information was extended with additional information to cover all relevant metadata information of different satellite data formats on the most scientific interesting missions, like LANDSAT, MODIS, S-NPP VIIRS, RapidEye and ESA Sentinel-2A and Sentinel-1A. ISO 19139-2 and its extension were incorporated in the spatial catalog Geonetwork Opensource by implementing new schema plugins. The schema plugin developed for ISO 19139-2 was released for the Geonetwork community and can be downloaded at the Geonetwork github repository (https://github.com/geonetwork/schema-plugins). The implementation of this work provokes three important benefits. First, the automatic metadata extraction that replaces the manual metadata creation saves a lot of time for the creation of metadata. Second, the management of the most relevant remote sensor data via metadata inside a spatial catalog, has improved the organization of the data inside the institute. The extended search functionality was improved to obtain the dataset, someone is looking for, easily and quickly. As a consequence a drastic reduction of data duplicates has occurred. Nowadays, thanks to the used catalog, researchers can interactively browse through spatial data without the need in additional GIS software. Last but not least, the produced results are beneficial for the entire Geonetwork community. The integration of the standard ISO 19139-2 as a new schema plugin in this widely used catalog and providing this plugin publicly available expand the application area of the catalog to cover the EO application field.

For the future development it is planned to expand the metadata format and the schema plugin for other type of sensor data. Inspired by the mission of Sentinel-2 launched in March 2014, in the near future a new extension of ISO 19115-2 will be elaborated in order to satisfy all requirements of EO metadata. Moreover, on the basis of new projects starting in field of environmental monitoring, a new metadata schema dedicated for ground sensor data will be finalized.

Acknowledgments We would like to thank to Marcello Pettita for providing assistance during the writing process.

Appendix A: ISO 19115-2 standard for describing imagery and gridded data

Most catalogs use a standardized format (XML) for structuring their metadata. There are different ISO standards for organizing metadata in a structured representation. The standard ISO 19115 provides an abstract and logical model

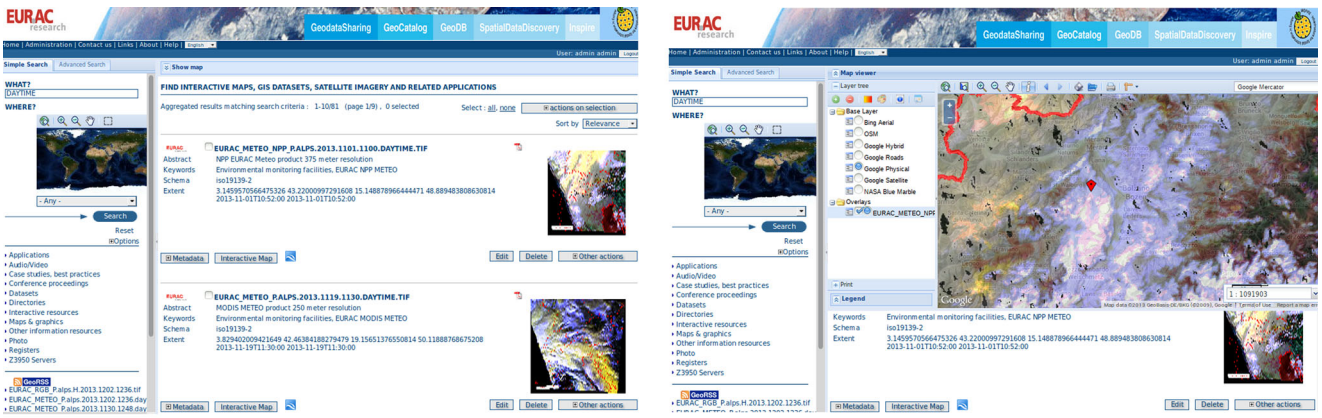

Fig. 7 Managing EO-data with Geonetwork OpenSource

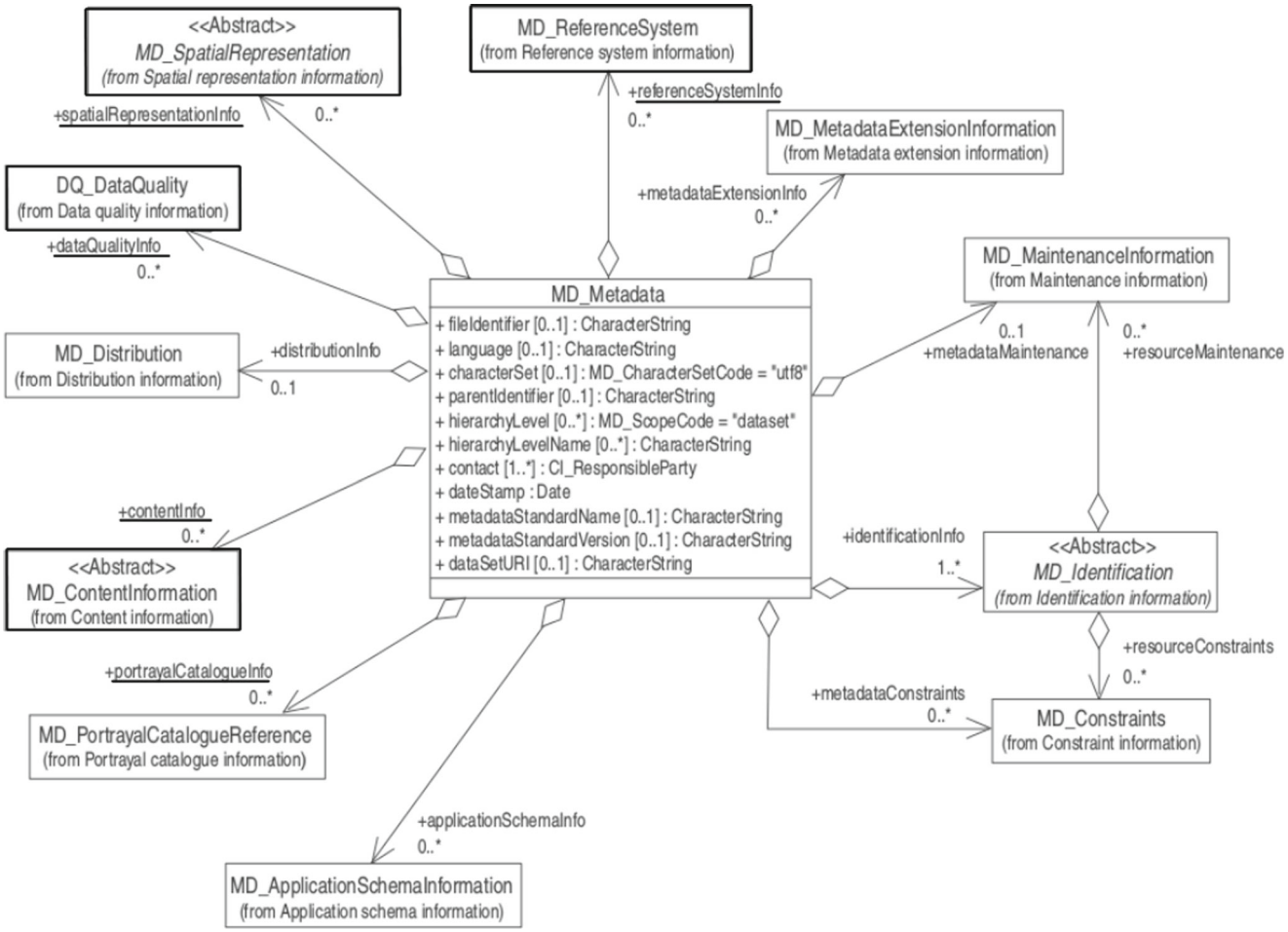

Fig. 8 UML Schema for Geospatial Metadata (ISO-19115 (National Coastal Data Development Center et al. 2012a)). Solid marked rectangles are extended in ISO-19115-2 and a new element MI AcqusitionInformation is introduced

for the organization of geospatial metadata. Because of its international acceptance, this standard has become a part of the "OpenGIS Specification" as the abstract model for the management of metadata. ISO 19115 defines the metadata to record information about the identification of geospatial data sets, possible approaches for distributing the data, details about the quality, geographic facets about the coordinate reference system and the coverage, temporal information about the date of acquisition, the owner of the data. The metadata schema is presented in form of UML class diagrams. Figure 8 presents the package for ISO 19115 containing all metadata child classes of the root class *MD Metadata*. The prefix *MD* stays as abbreviation for metadata.

However, this standard can not fulfill a big part of the requirements for imagery and gridded data. In order to cover these missing information in a standardized format, there was published the extension for gridded and imagery data (ISO 19115-2) in the year 2009. That new standard details, how a dataset was acquired, what kind of instrument has been used to produce the data, what kind of algorithm was used to process the data and remarks about the quality of a dataset. The name of the schema became MI-Metadata and all newly introduced elements start with the prefix *MI*. The new schema is an extension of ISO-19115 and this enhancement was applied on the root class and to all classes that are marked with a solid rectangle in Fig. 8 (MD ContentInformation, MD SpatialRepresentation, MD ReferenceSystem). Also a new subclass MI AcquistionInformation was introduced.

MD ContentInformation is extended with new child elements to describe the content of a coverage. The element MD Band (→ MI Band) was extended with additional attributes for specifying properties of individual wavelength bands. The element MI RangeElementDescription was added to provide identification of the range of elements used in a coverage dataset. Element MD ImageDescription is extended to include MI RangeElementDescription

in MI Image Description and MD CoverageDescription is also extended to include MI RangeElementDescription in MI Coverage Description.

The spatial representation package MD SpatialRepresentation is extended to include check point information to further specify georectification details, from MI GCP, in MI Georectified. MD Georeferenceable is extended to include additional information that can be used to geolocate the data, from MI GeolocationInformation, in MI Georeferenceable.

Moreover a new section for describing the acquisition process was added. MI AcquisitionInformation includes several other new packages. Figure 9 illustrates the acquisition package with the main class MI AcquisitionInformation that includes: the class MI Operation provides information of the overall data gathering program. The class MI Platform provides information about the platform from which the data were taken. Class MI Instrument provides designation of the measuring instruments used to acquire the data. Class MI Objective describes the characteristics and geometry of the intended object to be observed. The class MI Requirement defines the requirements used to derive the acquisition plan. The class MI Plan details the implementation and class MI Event expresses a significant even that occurred. The class MI PlatformPass identifies a particular pass made by the platform during data acquisition.

The class MD ReferenceSystem was extended with the class MD CRS including additional information about the projection, the ellipsoid, the datum, some details about the projection parameters (zone, scale factors, oblique line azimuth and oblique line point) and the ellipsoid parameters.

Appendix B: Acronymus table

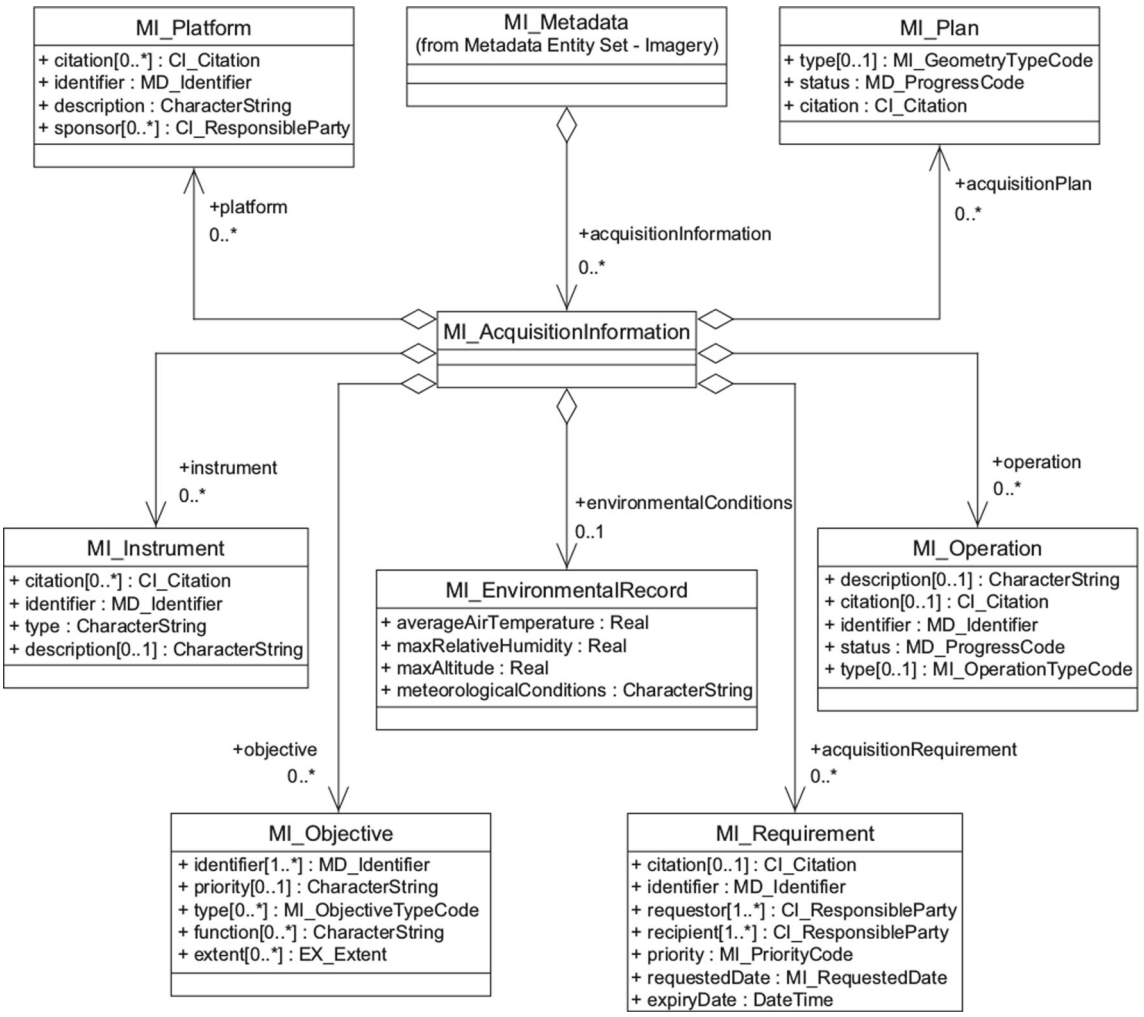

Fig. 9 UML Schema for Imagery and Gridded Metadata (ISO-19115-2 (National Coastal Data Development Center et al. 2012b))

Table 2 Acronymus table

SDI	Spatial Data Infrastructure
WMS	Web Mapping Service
EO	Earth Observation
SDI	Spatial Data Infrastructure
XML	eXtentable Markup Language
XSLT	eXtensible Stylesheet Language Transformations
GIS	Geographic Information Systems
MODIS	Moderate Resolution Imaging Spectroradiometer
VIIRS	Visible Infrared Imaging Radiometer Suite
COSMO-SkyMed	COnstellation of small Satellites for the Mediterranean basin Observation
ASAR	Advanced Synthetic Aperture Radar
NRT	Near Real Time
SDI	Spatial Data Infrastructure
SDI	Spatial Data Infrastructure

References

Dubitzky W, Wolkenhauer O, Yokota H (2011) Rule-Based Methods in Encyclopedia of Systems Biology. Springer

Giles HHCL, Manavoglu E, Zha H, Zhang Z, Fox EA (2003) Automatic document metadata extraction using support vector machines. In: JCDL Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries. IEEE Computer Society, pp 37–48

Greenberg J, Spurgin K, Crystal A (2005) Final report for the amega (automatic metadata generation applications) project. School of Information and Library Science. University of North Carolina, Chapel Hill

Litwin L, Rossa M (2011) Geoinformation metadata in INSPIRE and SDI. Understanding, Editing. Puslishing. Springer

National Coastal Data Development Center, National Oceanographic Data Center, N. O., Administration A (2012a) Iso 19115 2003 - geographic information - metadata, workbook. Technical report International Organization for Standardization, Geneva, Switzerland

National Coastal Data Development Center, National Oceanographic Data Center, N. O., Administration A (2012b) Workbook: Geographic information metadata xml schema implementation part 2: Extensions for imagery and gridded data. picture International Organization for Standardization, Geneva, Switzerland

Olfat H, Rajabifard A, Kalantari M (2010) Automatic spatial metadata update: a new approach. FIG Congress 2010 - Facing the Challenges - Building the Capacity

Ostensen O, Danko DM (2005) World Spatial Metadata Standards: Global Spatial Metadata Activities in the ISO/TC211 Geographic Information Domain. Elsevier

Various (2013) Geonetwork developer manual. online

Wilson M (2008) Metadata - Describing geospatial data. SDI cookbook