

ONTOLOGY BASED SEMANTIC METADATA FOR IMAGERY AND GRIDDED DATA

CHEN Xu*, ZHU Xinyan, DU Daosheng

the State Key Lab of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University 129,
Luoyu Rd., Wuhan, Hubei, China, 430079

WG II/6

KEY WORDS: Satellite Remote Sensing, Interoperability, Metadata, Semantics, Image Understanding, Ontology, Geo-information Management

ABSTRACT:

Nowadays, huge volumes of imagery and gridded data are accessible. However they vary in formats and are stored at various organizations leading to problems of data discovery, data interoperability and usability. ISO19115, Content Standard for Digital Geospatial Metadata and some other norms don't provide enough provision for imagery and gridded data. UML and XML schema of metadata also don't provide semantic description of the data content within the dataset. So we provide the additional structure to describe the derivation of imagery and gridded data. This structure is intended to augment the existing metadata standard described in ISO 19115. We focus on identification information, data quality information, spatial representation information, content information, acquisition Information in ISO 19115, we extend these dataset packages, in order to describe the metadata of imagery and gridded data during the cycle of the image product. OWL permits a much richer table of semantics as well as a more flexible definition of classes and their attributes when compared to UML and XML schema. We propose the mapping rule between UML and OWL, summarize the similarities between UML and OWL concepts. Then we explain in detail the conventions we have adopted to translate the metadata UML model for imagery and gridded data into an OWL ontology, including naming conventions, data types, and restriction conditions between the two concepts. Finally, we store imagery and gridded data metadata ontologies and individuals in a semantic registry prototype, to demonstrate how to fully utilizing the semantic information implicitly embedded in metadata.

1. INTRODUCTION

In geospatial domain, every day, large amounts of data are produced by different providers, archived at various locations and distributed in many different formats. This variety of formats leads to data interoperability and data usability problems faced by the researchers and other users. Also, the datasets are distributed and stored by various organizations making the task of locating and retrieving the relevant datasets very complex. There is a vital need of an efficient mechanism for discovery of required datasets. Geospatial metadata norm is developed to facilitate data sharing among geospatial information communities, such as ISO19115 (ISO 19115, 2003), the Content Standard for Digital Geospatial Metadata (CSDGM) (CSDGM, 1998). But these norms don't provide enough provision for imagery and gridded data and UML and XML schema of metadata don't provide semantic description of the data content within the dataset.

In this paper, we provide the additional structure to describe the derivation of imagery and gridded data. This structure is intended to augment the existing Metadata standard described in ISO 19115. We propose the mapping rule between UML and OWL and build the whole OWL metadata ontology for imagery and gridded data based on the rules. First, we analyze ISO19115, CSDGM and some kinds of imagery and gridded data formats, such as Dimap of spot. FGDC Content Standard for Digital Geospatial is very complex with 334 different elements, 119 of which exist only to contain other elements making this standard difficult to use (Parekh, 2004). We choose

the ISO 19115 to extend for imagery and gridded data, because it is the most general standard around the globe, and ISO19115-2 draft (ISO19115-2, 2007) is the foundation of our metadata model. In ISO19115, metadata for geographic data is presented in UML Packages. It consists of fifteen dataset packages. We will focus on identification information, data quality information, spatial representation information, content information, acquisition information, requirements information, extent information, citation and responsible party information, we extend these dataset packages, in order to describe the metadata of imagery and gridded data. Identification information contains information to uniquely identify the data. Additional identification metadata specified in this part includes entities to describe references that apply to the data and entities to identify the components used to acquire the data. The data quality package is defined as a container of a general assessment of the quality of the dataset. In addition, the package contains information about the sources and production processes used in producing a dataset, which is of particular importance for imagery and gridded data. The spatial representation package contains information concerning the mechanisms used to represent spatial information. Content information package expands on the information content that can be used to describe the content of a coverage dataset. Acquisition Information package provides details specific to the acquisition of imagery and gridded data. Requirements Information package provides details specific to the tasking and planning associated with the collection of imagery and gridded data. Extent information package defines datatypes that are used in other metadata sections and throughout the ISO 191xx

* CHEN Xu, Ph. D Candidate; email: manangelchenxu@yahoo.com.cn; phone:13638618595

suite of standards, this package is an aggregate of metadata elements that describe the spatial and temporal extent of the referring entity. Citation and responsible party information defines datatypes that standardize the method for citing a resource as well as information about the party responsible for a resource. All of these dataset packages are extended via the Unified Modeling Language (UML) for conceptualizing the underlying structure of metadata for imagery and gridded data.

Second, we prove that owl is to be a better choice for encoding the metadata for imagery and gridded data, because it permits a much richer table of semantics as well as a more flexible definition of classes and their attributes when compared to UML and XML schema. We propose the mapping rule between UML and OWL, UML class can be expressed in a similar concept using owl:Class, UML association can be mapped into owl:ObjectProperty and UML attribute can be expressed as owl:DatatypeProperty. On other side, there are some concepts in OWL which cannot be mapped as straight forward into UML. We summarize the similarities between UML and OWL concepts, and also indicate the most significant dissimilarities. Then, we explain in detail the conventions we have adopted to translate the metadata UML model for imagery and gridded data into OWL ontology, including naming conventions, data types, and restriction conditions between the two concepts. After we have developed a mapping approach for imagery and gridded metadata schema, we translate the metadata model from its conceptualization in UML into an ontology using the Web Ontology Language, OWL. We use the protégé to implement metadata ontologies for imagery and gridded data.

Finally, we develop a semantic metadata registry prototype. The objective of this prototype is to provide a metadata paradigm that is semantically rich and capable of facilitating content based discovery of imagery and gridded datasets to the end users.

Metadata, as an important tool to facilitate the discovering and sharing of geographic data, faces the problem of semantic heterogeneities because various metadata are published according to different metadata standards such as ISO 19115, CSDGM and so on. These metadata specifications provide different framework for the production of metadata which is a factor leading to semantic heterogeneities. To solve the semantic heterogeneities among metadata, ontology-based metadata has been studied by several researchers such as (Parekh, 2004), (Islam, 2004). Islam A. et al are developing a metadata ontology based on FGDC metadata standard making it very complex and difficult to use as compared to our ontology. There are other on-going projects in using Semantic Web technologies to improve data discovery, usability and interoperability. As a part of Semantic Web for Earth and Environmental Terminology (SWEET) (Raskin, 2003) project at NASA, they have developed several domain ontologies to describe earth science data and knowledge. Their motivation is to improve the discovery of NASA information and data products. However, the formal metadata norm for imagery and gridded data does not publish. Moreover, none of them strive to develop to a semantic metadata standard for imagery and gridded data. Our use of Web Ontology Language (OWL) provides more semantic power to the metadata for imagery and gridded data and also makes the semantic metadata files available to the next generation Semantic Web (Berners-Lee, 2001). Our study facilitates ontology based querying for datasets compared to keyword searches for XML Schema metadata.

2. EXTENSIOINS FOR IMAGERY AND GRIDDED DATA

2.1 The package structure of ISO19115

The objective of ISO19115 is to provide a general-purpose metadata for describing digital geographic data. This International Standard defines metadata elements, provides a schema and establishes a common set of metadata terminology, definitions, and extension procedures. More detailed metadata for geographic datatypes and geographic services are defined in other ISO 191xx suite of standards and user extensions.

ISO19115 standard includes fifteen packages, these packages provide information about the identification, the extent, the data quality, the spatial and temporal schema, spatial reference, distribution of digital geographic data and so on. Figure 1 provides the relationship between these metadata packages. This International Standard defines: mandatory and conditional metadata sections, metadata entities, and metadata elements, the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data); optional metadata elements – to allow for a more extensive standard description of geographic data, if required; a method for extending metadata to fit specialized needs.

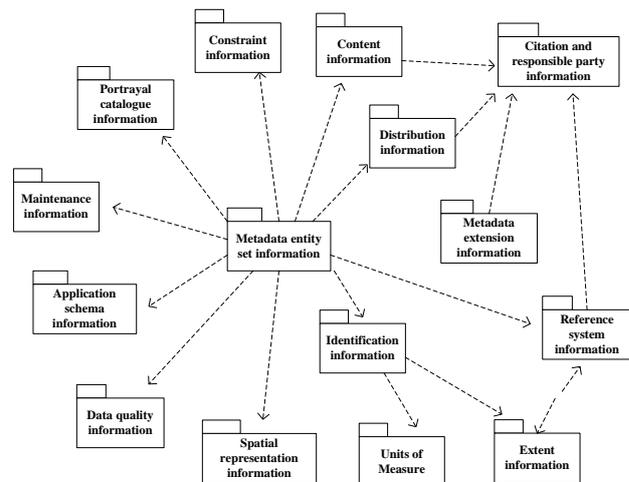


Figure 1. Metadata packages in ISO19115

Metadata entity set information defines the metadata about a resource and establishes hierarchy. Identification information is the basic information needed to uniquely identify a resource. Constraint information is about legal and security restrictions on access and use of data. Data quality information is used to describes data quality and testing. Lineage information records sources and production processes. Maintenance information describes maintenance and update practices, it contains information about the scope and frequency of updating data. Spatial representation information describes mechanism used to represent spatial information (grid or vector). Reference system information describes spatial and temporal reference system used in a dataset. Content information contains information identifying the feature catalogue used and information describing the content of a coverage dataset. Portrayal catalogue information contains information identifying the

portrayal catalogue used. It contains distributor information and methods of access. Metadata extension information is about user specified extensions. Application schema information contains information about the application schema used to build a dataset.

ISO 19115 includes two metadata datatype packages: extent information and citation and responsible party information. Extent information is an aggregate of the metadata elements that describe the spatial and temporal extent of the referring entity. It contains information about the geographic, temporal and the vertical extent of the referring entity. Citation and responsible party information provides a standardized method for citing a resource (dataset, feature, source, publication, etc.), as well as information about the party responsible for a resource.

2.2 How to extend the ISO19115 for imagery and gridded data

The ISO 19115 Metadata model does provide some provisions for imagery and gridded data, but imagery and gridded data have their own geographic features which need to be described in metadata. For example the production process needs to be documented in order to maintain quality control over the end products. The object of ISO 19115-2 draft is to provide the additional structure to more extensively describe the derivation of geographic imagery and gridded data, but now it is not a final norm and it is lack of some import parameters. We analyse the ISO 19115-2, modify some entities in the draft, and extend the existing geographic metadata standard by defining the schema required for describing imagery and gridded data, we focus on identifier information, data quality information, content information, acquisition information. Then we build the metadata model upon it. Figure 2 illustrates the relationships among the packages described in this part of ISO 19115 and the relevant packages specified in own metadata model.

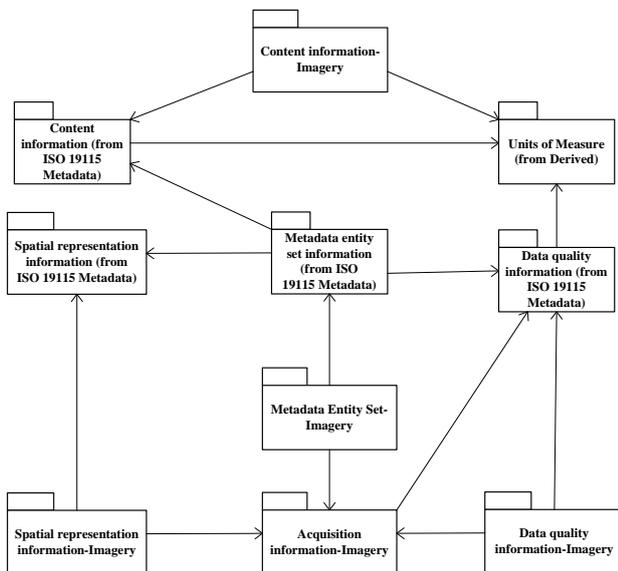


Figure 2. Metadata packages for imagery and gridded data

Table 1 contains the list of package identifiers for the classes used in ISO191XX serials. In this study, we inherit this convention.

Identifier	Information Type	Standard
CI	Citaaion	ISO 19115
DQ	Data Quality	ISO 19115
EX	Extent	ISO 19115
GM	Gemometry	ISO 19107
LE	Lineage Extend	ISO 19115-2
LI	Lineage	ISO 19115-2
MD	Metadata	ISO 19115
MI	Metadata for Imagery	ISO 19115-2
MX	Metadata-XML schema	ISO/TS19139
QE	Data quality Extended	ISO 19115-2

Table1. UML Package Identifiers

Now, we introduce the extensioins for imagery and gridded data. We add some entities into different packages, such as MI_Metadata, MX_DataFile, MI_AcquisitionInformation, QE_CoverageResult, LE_Source, LE_ProcessStep, LE_Processing, LE_ProcessStepReport, LE_Algorithm and so on. Metadata entity set is the root package of the metadata, an additional package of metadata is specified. MI_Metadata is a specified subclass of MD_Metadata. It is used to provide information describing imagery and gridded data. The extension provided through MI_Metadata adds an association to the MI_AcquisitionInformation class. Some others package is introduced in details in following.

2.2.1 Data quality information: Data quality information is very important for imagery and gridded data, because the production process needs to be documented in order to maintain quality control over the end products. In addition, metadata about the geometry of the measuring process and the properties of the measuring equipment needs to be retained with the raw data in order to support the production process. But the original data quality package is defined in ISO 19115 as a container of a general assessment of the quality of the dataset. We add the following classes to those specified in ISO 19115:

QE_CoverageResult is a specified subclass of DQ_Result and aggregates information required to report data quality for a coverage. From ISO 19115, the classes MD_SpatialRepresentation, MD_CoverageDescription, and MD_Format are aggregated as a description of the coverage data quality result. In addition, an association with MX_DataFile (defined in ISO/TS 19139) provides a means to identify a complete report of the quality of the coverage. QE_Usability is a specified subclass of DQ_Element used to provide user specific quality information about a dataset's suitability for a particular application. LE_ProcessStep is a specified subclass of LI_ProcessStep and contains additional information on the history of the algorithms used and processing performed to produce the data. LE_ProcessStep aggregates the following entities: LE_Processing describes the procedure (such as software used, parameters, and processing documentation) by which the algorithm was applied to generate the data from the source data. LE_Processing aggregates LE_Algorithm, which describes the methodology used to derive the data from the source data. LE_ProcessStepReport identifies external information describing the processing of the

data.LE_Source is a specified subclass of LI_Source and describes the output of a process step.

2.2.2 Spatial representation information: The spatial representation package contains information concerning the mechanisms used to represent spatial information. MI_GeoreferencingDescription is a specified subclass of MD_Georeferenceable that contains addition information used to support georectification of the data. MI_GeoreferencingDescription is an aggregation of the following entities:

MI_Georectified is a specified subclass of MD_Georectified that contains check point information to further specify georectification details of the imagery or gridded data. MI_Georectified aggregates MI_GCP.MI_Georeferenceable is a specified subclass of MD_Georeferenceable that includes additional information that can be used to geolocate the data. MI_Georeferenceable aggregates MI_GeolocationInformation.

2.2.3 Content information: The package Content Information describes the content of a coverage dataset. Our model expands it like this:

MI_Band is a specified subclass of MD_Band defining additional attributes for specifying properties of individual wavelength bands in an imagery and gridded dataset. MI_ImageDescription is a specified subclass of MD_ImageDescription used to aggregate MI_RangeElementDescription. MI_CoverageDescription is a specified subclass of MD_CoverageDescription used to aggregate MI_RangeElementDescription. MI_RangeElementDescription provides identification of the range elements used in a coverage dataset.

2.2.4 Acquisition Information: This package is particular to imagery and gridded metadata and provides details specific to the acquisition of imagery and gridded data. MI_AcquisitionInformation is an aggregate of the following entities:

MI_Instrument designate the measuring instruments used to acquire the data. MI_Operation designate the overall data gathering program to which the data contribute. MI_Platform, designate the platform from which the data were taken. MI_Objective describes the characteristics and geometry of the intended object to be observed. MI_Requirement describes the user requirements used to derive the acquisition plan. MI_Plan illuminate the acquisition plan that was implemented to acquire the data. Two additional classes are required to provide information on the acquisition of the data. These are: MI_Event and MI_PlatformPass. MI_Event describes a significant event that occurred during data acquisition. An event can be associated with an operation, objective, or platform pass. MI_PlatformPass identifies a particular pass made by the platform during data acquisition. A platform pass is used to provide supporting identifying information for an event and for data acquisition of a particular objective.

When we finish the extension of ISO19115, we need to implement this metadata model in order to use it in geospatial system.

3. SEMANTIC METADATA FOR IMAGERY AND GRIDDED DATA

We extend the ISO19115 norm to support the description of imagery and gridded data, which is implemented via the Unified Modeling Language (UML) for conceptualizing the underlying structure of the metadata model. Neither XML Schema, nor RDF, nor RDF Schema fully satisfies the specific needs of a structure to resolve semantic problems that one can easily foresee emerging when using the ISO 19115:2003 metadata norm (Islam, 2004). The use of the Web Ontology Language (OWL), built on XML, proves to be a better choice for encoding the metadata model because it permits a much richer table of semantics as well as a more flexible definition of classes and their attributes when compared to XML schema, RDF and RDF Schema. In this paper suggest a number of mapping approaches that we employed during the mapping effort while trying to overcome some of the incompatibilities that exist between UML and OWL. Then we create an ontology for imagery and gridded metadata that implements the metadata model in a machine understandable format.

3.1 OWL and UML

Ontology is designed for the purpose of enabling knowledge sharing and reuse .Ontologies can be applied in many ways such as for detecting inconsistencies, performing validations, extending metadata, ensuring interoperability, and for the validation of information integrity and hierarchy (Fensel 2003). OWL (Smith, 2004) is the most recent development in standard ontology languages, endorsed by the World Wide Web Consortium (W3C) to promote the Semantic Web vision. An OWL ontology may include descriptions of classes, properties and their instances. The UML representation of the metadata model is optimized to guide efficient implementation. Thus, a goal is to use as few classes as possible. Ontology design, on the contrary, aims to provide extensive semantic definitions of concepts.

3.2 Mapping of UML model and OWL ontology

Mapping rules for transforming a UML model into a DAML ontology are described in (Baclawski, Kokar et al., 2001) and can easily be transferred to OWL. Although their purposes are different, UML and OWL share many characteristics. For example, UML class can be expressed in a similar concept using owl:Class, UML association can be mapped into owl:ObjectProperty and UML attribute can be expressed as owl:DatatypeProperty. However, it is important to realize that the conceptualization of the metadata model (which is provided in UML) and its conversion into a machine readable framework is not straightforward but fraught with shortcomings and potential pitfalls. There are some concepts in OWL which cannot be mapped as straight forward into UML. For example, there is no equivalent concept in UML to express the owl:unionOf concept of OWL. A summary of the similarities between UML and OWL concepts has presented in Table 2.

UML concept	OWL concept
Package	Ontology
Class	Class
Attributes	Data property
Associations	Object property
Generalization Relation	Hierarchy
Data Type (user-definable)	Data Type
DataType (primitive pre-	xsd: type

defined)	
Object	Instance
Multiplicity Constrains	Restrictions

Table 2. Similarities between UML and OWL concepts

Before we start to translate (or map) the metadata UML abstract model into an OWL-based ontology, it is worthwhile to further discuss some of the dissimilarities, such as incompatibility of properties, monotonic worlds, modularization, generalization, abstract class and so on. For example, a metadata element may not be globally unique as other entities can contain the same element name. This could be a problem if we are using name as identifiers (e.g. rdf:ID), which should be unique throughout the namespace. We use the following rules to resolve this problem. a. If duplicate elements have the same range, only one element is declared without defining its domain. A class that uses this property could refer to it by creating a restriction. b. If duplicate elements have different range, create a new element using its short name. c. If duplicate element is a code list element, an underscore (“_”) is added before its name. As we have only pairs of duplicate elements, this approach can resolve the conflict successfully. In (Islam, 2004), they explain in detail the conventions they have adopted to translate the ISO 19115:2003 UML model into an OWL ontology, we obey their rules in this study.

3.3 Metadata Ontology for Imagery and Gridded Data

After we give mapping rules between OWL and UML, we build metadata ontology for imagery and gridded data. The following text shows a Description fragment from this metadata ontology.

```
<owl:Class rdf:ID="MI_AcquisitionInformation">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:maxCardinality
rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
      >1</owl:maxCardinality>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="instrumentIdentification"/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
  <iso19115:_shortName
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >PltfrmInstId</iso19115:_shortName>
  <rdfs:label
xml:lang="en">MI_AcquisitionInformation</rdfs:label>
```

We can see MI_AcquisitionInformation entity is converted to owl: Class, owl: Cardinality is used to mapping the max occurrence of referencing object. The entities that compose the MI_AcquisitionInformation entity are expressed by owl:ObjectProperty. We build the whole OWL metadata ontology for imagery and gridded data. It is being made available to the general public at http://www.liesmars.wtusm.edu.cn/prof_web/zhuxinyan/my/ontology/19115-ex.

4. CONCLUSION

In this paper, we discussed our metadata model for imagery and gridded data, and explained in detail the conventions we have adopted to translate the metadata UML model for imagery and gridded data into an OWL ontology. At last, we build the

metadata ontology for imagery and gridded data. We can publish imagery and gridded data metadata ontologies and individuals in a semantic registry prototype that we study in the (Chen, 2007). Each imagery and gridded dataset that is registered with our OWL ontologies has content based semantic description. This semantic description is independent of the dataset format and is generated using the geospatial domain specific ontologies. This approach allows the end users of data to search for relevant datasets based on their semantic content rather than just simple keywords.

REFERENCE

Baclawski K., Kokar M K., Kogut P A., Hart L., Smith J., Letkowski J., and Emery P., 2002. Extending the Unified Modeling Language for ontology development. *Software System Model* 1: 1-15.

Berners-Lee T., Hendler J., Lassila O., 2001. The Semantic Web, *Scientific American*.

Chen X., Zhu X.Y., Zhang X., 2007. Geospatial Data and Services Semantic Share based on ebXML Registry, *The Second International Conference on Space Information Technology*.

Federal Geographic Data Committee, 1998. Content standard for digital geospatial Metadata, Federal Geographic Data Committee.

Fensel D., 2003. *Spinning the semantic Web: bringing the World Wide Web to its full potential*. Cambridge, Mass., MIT Press.

Islam, A., Bermudez, L.E., and Piasecki, M., 2004. Ontology for Geographic Information - Metadata (ISO 19115), <http://loki.cae.drexel.edu/~wbs/ontology/>.

ISO/TC-211 (2003): ISO 19115:2003. Geographic information - Metadata, International Organization for Standardization.

ISO/TC 211 19115-2 Geographic information - Metadata - Part 2(2007): Extensions for imagery and gridded data.

Parekh, V. et al, 2004. Ontology based Semantic Metadata for Geoscience Data, *Proceedings of the International Conference on Information and Knowledge Engineering, IKE'04*, Las Vegas, Nevada, USA, CSREA Press, pp.485-490.

Probst, F., Gibotti, F., Pazos, A., 2004: Connecting ISO and OGC Models to the Semantic Web, *3rd International Conference on Geographic Information Science*.

Raskin R., Pan M., 2003. Semantic Web for Earth and Environmental Terminology (SWEET), *Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, Florida, USA,

Smith, M., Wetly C., Deborah L., 2004. OWL Web Ontology Language Guide, <http://www.w3.org/TR/owl-guide/>

ACKNOWLEDGEMENTS

The work described in this paper is supported by Western Mapping Project of China, National Natural Science Foundation of China (No.40701156) and Hubei Natural

Science Fund for Innovation groups projects (No: 2006ABC010).