ONTOLOGY BASED APPROACH FOR WATER RELATED INFORMATION SYSTEM FOR MEKONG DELTA, VIETNAM

Tran Thai Binh[1, 2], Verena Klinger[2] and Klaus Greve[3]

[1]GIS and Remote Sensing Research Center, Hochiminh city Institute of Resources Geography, Vietnam
Email: ttbinh@vast-hcm.ac.vn / binh.tranthai@dlr.de
[2]Remote Sensing Date Center, German Aerospace Center, Germany
Email: verena.klinger@dlr.de
[3]Geography Department, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
Email: klaus.greve@uni-bonn.de

ABSTRACT

The German-Vietnamese water-related information system for Mekong Delta (WISDOM) project is a multidisciplinary project associated with the principle of IWRM (Integrated Water Resources Management) which is defined as a multidisciplinary approach in order to manage water and related resources for sustainable development. WISDOM develops and implements an innovative water-related information system containing all the outcomes and results of the multidisciplinary research involved in the project. However, the collected data use several models and terminology to describe the same real-world object or phenomenon which causes semantic heterogeneity. As a result, finding and accessing the appropriate data or information is not straightforward. A database with a cross relation structure however, is not able to manage semantic heterogeneity issues of collected data from research fields. A forward looking solution of providing all relevant data precisely for a specific query in the WISDOM information system is to resolve the semantic heterogeneity of data.

This paper presents a new approach applying ontology for data discovery and retrieval for the WISDOM information system. Ontology is applied to resolve the constraints of existing structure on describing the relationships between datasets and thematic reference groups. Within this new approach, all datasets are described by linkages to superior subjects so that relevant data sets of multidisciplinary fields are provided by only one search. All in all, the ontology approach facilitates user search for data being more precise and suitable for their demands. As a characteristic of ontology, this approach ensures for transferability and scalability to other domains.

1. INTRODUCTION

The WISDOM project (www.wisdom.eoc.dlr.de) is a multidisciplinary project associated with the principle of IWRM (Integrated Water Resources Management) which is defined as "*a process, which promotes the coordinated development and management of water, land and related resources in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems"* (Global Water Partnership, 2000). The project focuses on development and implementation of an innovative water-related information system (IS) containing all the outcomes and results of the different research disciplines involved in the project (WISDOM, 2011). The WISDOM IS is web based data infrastructure applying internet infrastructure and related technologies on information from a variety sources like sensors networks, field based surveys, census data and earth observation (WISDOM, 2011). It consists of huge amount of data from several research fields, from geographic data and satellite imageries, sensor data that show water quality, water level to statistical data, reports and literatures in the field of water knowledge, livelihoods and knowledge management. The collected data are heterogeneous because of different formats, different scales and areas of interest and different disciplines.

This paper focuses on the drawbacks of the cross-related structural database which was used to manage the thematic reference aspects, and proposes an ontology based approach to overcome these issues.

2. THEMATIC REFERENCE SCHEMA IN WISDOM

To retrieve and use appropriate data from the WISDOM IS, the user needs information about the stored data. These metadata are managed and generated from information from data provided by the data producers. The WISDOM data management model organizes dataset attributes into aspects. The aspects are divided into two groups to describe datasets in detail and to define how data is stored, i.e. data aspects and reference aspects (Gebhardt et al., 2010).

Within the WISDOM IS, data aspects act as metadata which advances data query and data distribution algorithms, e.g. spatial datasets can be searched by ISO19115 and ISO 19139 metadata using OGC Geoservices (OGC, 2012) such as the Geonetwork catalogue system (see more details in (WISDOM, 2012)). Furthermore, a dataset can be retrieved as a WMS layer via common web or desktop clients such as OpenLayers, Gaia or ESRI ArcMap (Gebhardt et al., 2010). In addition, the reference aspects are designed to allow users to explore data using efficient search options by thematic, geographic and temporal search variables. The reference aspects are the following: (1) The **Spatial reference aspect** presents a hierarchical structure of administrative areas, in which the observations or survey data are collected, according to administrative level, i.e. country, region, province, district and commune; (2) The **Temporal references aspect** contains the instants value of time describing the valid period of datasets; (3) The **Thematic references aspect** is a list of themes which is organized as hierarchical groupings, which enables to access to hydrologic, environmental, or social data from general to more specific groups.

Product Group ID	Product group name	Reference theme Id	Reference theme name	Reference theme level
16	River network	1	Environment	1
16	River network	2	Infrastructure	1
16	River network	8	Hydrology	1
16	River network	10	Transportation	2
17	Water mask	14	Environment	1
17	Water mask	28	Hydrology	2
17	Water mask	29	Water level	3

Table 1: Examples of “product-theme” entity relation model in WISDOM IS
(Source: (Gebhardt et al., 2010))

Implicitly, datasets which are mapped to one level in thematic hierarchy are also belongs to all the parent levels along the branch. Thus, datasets are assigned to its respective themes at different levels to speed up the data access process. To be able to double register one dataset, a n:m relationship between thematic reference and product group was set up (Klinger et al., 2010). Product groups are groups of datasets which are created by the same processing method and/or describe the same object in the real world. They are described as sub classes of thematic reference groups and are very close to the meaning to the term “dataset”, e.g. “watermask from optical sensors”, “water chemical substance”, “rainfall” and “soil moisture”, etc. As shown in figure 1, firstly, datasets are tagged by product group using 1:1 relationship, then each product group is mapped to a thematic reference group with a n:m relationship (Klinger et al., 2010). Table 1 shows an example on how a data assigned to thematic reference classes. A “watermask” belongs to “Environment” at the highest level, “Hydrology” in the next level and, finally to ‘‘Water level” at the lowest level. Also, spatial datasets are related to multiple themes within the same thematic level, e.g., the “River network” which belongs to “Environment”, and “Infrastructure” at the highest level. The thematic reference aspect adds thematic contextual information to a dataset using hierarchies. These relationships enable a meaningful search way through thematic groups.

Figure 1: The relationship between dataset and thematic reference via product group
(Source:(Klinger et al., 2010))

However, there are limitations of the model defined bellow:

(1) There are redundancies in the case of datasets related to different levels of the thematic reference schema. One thematic class connects to two or more higher level classes. One product group class links to two or more thematic classes (as shown in table 1). With this design, the extending of the thematic reference hierarchical structure may cause administration, maintenance and consistency problems (Gebhardt et al., 2010), because the complexity of the RDB increases when the relationships increase. In fact, extending that structure is complicated.

(2) Furthermore, in the case of IWRM when users need relevant data to analyse the influence between real-world objects, or legal documents, such as decisions and decrees of the different ministries for water-related field, the current WISDOM IS cannot provide necessary data in one search. The users have to search several times and change thematic criteria by themselves. The WISDOM IS is organised along thematic dimensions. In many cases, there are no relationships across the boundaries of thematic classes. The system cannot manage the relationships such as “relate”, “canObserve”, etc. For example, land cover relates to land use, or the number of farms relates to agriculture production.

The insufficiency of cross-related data structure can be solved by applying an ontology based approach in order to add semantic descriptions to the database. Ontology can describe the semantics of data in a machine readable way, it provides an appropriate way to manage relationships between datasets and reference aspects. In additions, there are some software programs which can visualize the structure of ontology so that extensions and maintenance of the ontology gets easier, e.g. Protégé software.

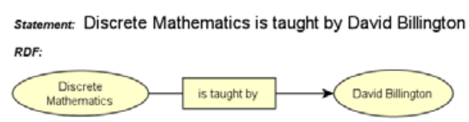

Figure 2: Example RDF
Source (Antoniou et al., 2008)

3. ONTOLOGY

Recently, ontology emerges as an appropriate tool to describe the meaning of data in a way that the computer can understand and apply meaningful data processing automatically. It is not only useful for sharing understanding, but also applied as the foundation to enhance data usage, to gain semantic interoperability, to develop advanced methods for representing and using complex metadata, correlating information, knowledge sharing and discovery.

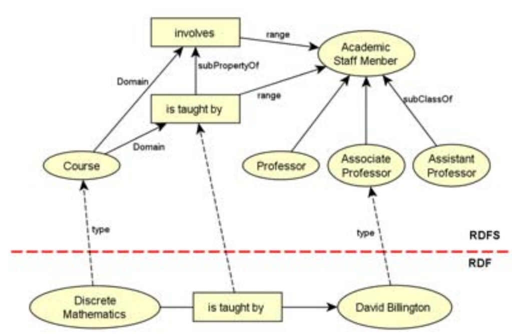

Figure 3: Example RDF and RDFs
Source (Antoniou et al., 2008)

To apply ontology, RDF (Resource Description Framework), was published in 1999 by W3C, can be used to describe objects (called resources) and their relationships on the web in a machine-understandable way (W3C, 2010a). In other words, metadata of data is available on the web. RDF uses a simple structure statement “Subject – predicate – object” to describe resources or to present the relation between resources in the structure “resource – property – resource/literal” (figure 2). RDF uses an extensible URI-based vocabulary with the XML syntax; hence exchange between different operating systems is easily possible. Any resource can be described with RDF statement (W3C, 2010b).

Since RDF is limited to the description of resources with classes, properties and values, further schema based on RDF was developed to extent the functionalities of RDF and broaden the potential application. **RDFs** (Resource Description Framework Schema) was developed based on RDF characteristics but it is extended to describe classes of resource and their properties (figure 3) such as class and subclass relations, “domain” and “range” restriction of properties. RDFs does not provide actual application-specific classes and properties, but the framework to describe it. Classes in RDFs are much like classes in object-oriented programming languages. This allows resources to be defined as instances of classes, and subclasses of classes. (W3C, 2010b)

RDF and RDFs were developed to provide basic capabilities for describing resources, but they specify fairly loose constraints on vocabularies. **OWL** (Web Ontology Language) was built on top of RDF and RDFs adding supplemental constraints to increase the accuracy of implementation of a given vocabulary. RDFs specifies fairly loose constraints on vocabularies. OWL adds supplemental constraints that increase the accuracy of implementations of a given vocabulary. Ontology also applies **reasoner** which is a program able to infer logical consequences from a set of asserted facts or axioms, i.e. Racer, Pellet, Fact++, Hermit etc. (Pan, 2005; Tsarkov et al., 2006; Sirin et al., 2007; Fahad et al., 2008). Applying a reasoner, the ontology constraints allow additional information to be inferred from the data, though it may not be explicitly represented in an ontology (for example if an individual Martin is in class Student, and the class Student is a subclass of the class Person, a reasoner will infer that Martin is a Person) (Powers, 2003). OWL uses XML syntax and is a recommendation of W3C for semantic web (W3C, 2012).

4. PROPOSED APPROACH

In general, data are created or collected in order to describe the status of objects in the real world. Thus, when searching for data, users actually want to get the information of an object at a certain time for a particular location. The objects in the real world, which can be described by datasets, are hereafter defined “observed object”. To overcome the issues mentioned above, an ontology is built. Beside the classes describing data aspects, spatial and temporal aspects, the proposed ontology contains classes representing concepts about observed objects related to water field. These concepts are applied instead of the thematic reference aspect.

Observed objects are presented in a hierarchical structure including properties that define how these objects relate to each other, e.g. the water mask has a relation to water level and the agriculture area relates to agriculture production. This hierarchy adopts concepts coming from the section agriculture from AGROVOC, dealing with the level of relations which are defined as broader-, narrower-, related term in AGROVOC. The broader terms present a more general concept than the narrower ones. The broader and narrower terms are in the

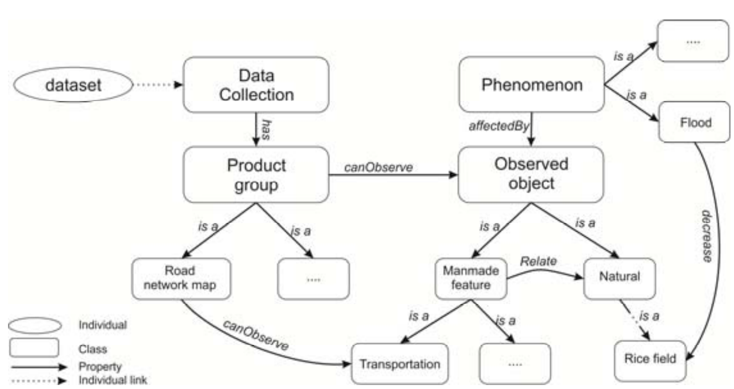

Figure 4: Abstract of relations of observed objects and product groups

same branch of the hierarchy. The related terms are in different branches but they related to each other. The higher level in the hierarchy presents more general concepts than the lower. For example, “natural” class has the sub class “land use and land cover” which contains “agriculture area”, “industrial area” etc. Thus, “natural” is a more general concept than “land use and land cover”. The more detailed concepts are “agriculture area”, “industrial area”. The concepts of AGROVOC thesaurus are widely applied by several communities (Sánchez-Alonso et al., 2007). With this approach, firstly, the observed object list and relationships of them are specified. The figure 5 shows the main classes under the observed object class. Then, the datasets with corresponding product group are assigned to proper observed object via the property “canObserve” as shown in figure 4.

Since the observed objects have interrelations themselves, i.e. the water mask has a relation to water level and the agriculture area relates to agriculture production. The links between product groups and observed object classes enable to infer the relations between datasets, even if these relations are not recorded in the database. For example, the datasets about the water level information can observe the status of the river under observation. A water mask dataset is a satellite imagery processing product in which all the water bodies are masked into a single layer presenting the distribution of surface water, so that it relates to water level. Therefore, it is inferred that water masks also relate to the river status. Figure 6 shows how the proposed ontology can discover appropriate datasets for user demands. With a certain query for an observed object (case 2 shown in figure 6), a list of product group is retrieved via the property “canObserve”. And finally, the datasets that belong to the defined product group classes are retrieved and provided to the user. Besides, the relevant datasets which belong to related observed object classes are also retrieved.

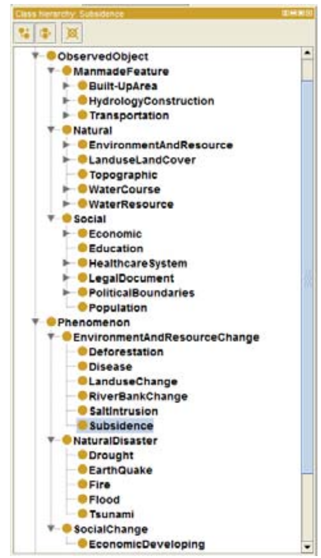

Figure 5: Classes hierarchy of proposed ontology, visualized by Protégé software

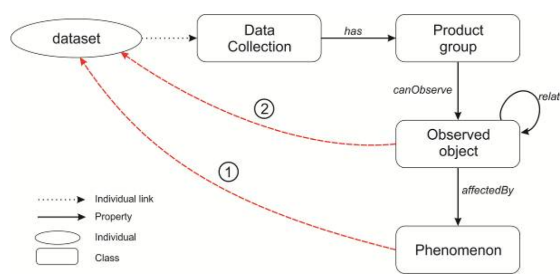

Figure 6: Abstract model for inference of datasets

Furthermore, we developed concepts which describe the phenomena in terms of influence with observed objects, for e.g. flood affects agriculture area. The properties such as change, destroy, decrease, and increase are used to describe the effect of phenomena to observed objects. These descriptions do not include the cause of phenomena. The list of phenomena is extracted from SWEET, Semantic Web for Earth and Environment Terminology (SWEET, 2012). The influence of phenomena on observed objects were acquired from several papers and definitions such as the study guide for disaster management of Schramm (Schramm et al., 1986) which list all the effects of common natural hazards. These relationships are used to infer which datasets relate to which phenomena and how datasets relate to each other in terms of their relations with observed objects. Figure 6 shows how the reasoners can infer datasets. With a certain phenomenon (case 1 in figure 6), observed objects are retrieved via “affectedBy” property, then product group via “canObserve” property. And finally, datasets which belong to the same product group classes are retrieved. By this way, user can discover and retrieve all relevant datasets for a phenomenon of interest with one search.

5. CONCLUSION

This paper presents an innovative approach applying ontology for data discovery and retrieval for the WISDOM information system. Ontology is applied to resolve the constraints of existing structure on describing the relationships between datasets and thematic reference groups. Within this approach, all datasets are described by linkages to observed object so that relevant data sets of multidisciplinary fields are provided by only one search. All in all, the ontology approach facilitates user search for data being more precise and suitable for their demands. As a characteristic of ontology, this approach ensures for transferability and scalability to other domains. It is easy to extend and maintain the ontology using ontology visualization software such as Protégé.

However, it is necessary to define the relationships between classes. That determines the results returned to user queries. In general, we can say that everything in the world relates to each other (Tobler, 1970) – which counts for user queries as well, the system retrieves every dataset because they are related to each other. On the other hand, some relevant data will not be retrieved if relationships are not or are inappropriate defined, which are described as the precision and the recall issues. Currently, this approach applied common knowledge adopted from pre-existing definitions. The assessment of the precision and recall should be considered for further works.

6. REFERENCES

Antoniou, G. and F. v. Harmelen (2008). "A Semantic Web Primer." *MIT Press, USA, 2004. ISBN: 0-262-01210-3*

Fahad, M., M. A. Qadir and S. A. Hussain (2008). "Evaluation of Ontologies and DL Reasoners." *Intelligent Information Processing IV, Volume 288/2008, Springer, 1571-5736 (Print) 1861-2288 (Online)*: 17-27.

Gebhardt, S., T. Wehrmann, V. Klinger, I. Schettler, J. Huth, C. Künzer and S. Dech (2010). "Improving Data Management and Dissemination in Web based Information System by Semantic Enrichment of Description Data Aspects." *Computers & Geosciences,* Vol. 36: 1362-1373.

Global Water Partnership (2000). http://www.gwp.org/.

Klinger, V., T. Wehrmann, S. Gebhardt and C. Künzer (2010). A Water-Related Web-based Information System for the Sustainable Development of the Mekong Delta. The mekong Delta System. Interdisciplinary Analyses of a River Delta. C. Künzer and F. G. Renaud. Springer Netherlands, Springer Environmental Science and Engineering. XV**:** 423-444.

OGC. (2012). "GeoServices Rest SWG " Retrieved 15th July, 2012, *from http://www.opengeospatial.org/projects/groups/gservrestswg*.

Pan, Z. (2005). "Benchmarking DL Reasoners Using Realistic Ontologies." *In Proc. of the OWL: Experiences and Directions Workshop, 2005.*

Powers, S. (2003). "Practical RDF." *O'Reilly 2003.*

Sánchez-Alonso, S. and M.-A. Sicilia (2007). Using an AGROVOC-based ontology for the description of learning resources on organic agriculture. Metadata and Semantics. M. A. Sicilia and M. Lytras, Springer**:** 481-492.

Schramm, D. and R. Dries (1986). Natural Hazards: Causes and Effects - Study Guide for Disaster Management. Disaster Management Center - University of Wisconsin-Madison.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur and Y. Katz (2007). "Pellet: A Practical OWL-DL Reasoner." *Journal of Web Semantics,* Vol. 5(2): 51-53.

SWEET. (2012). "Semantic Web for Earth and Environmental Terminology." Retrieved 14th June, 2012, *from http://sweet.jpl.nasa.gov/ontology/*.

Tobler, W. R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography,* Vol. 46: 234-240.

Tsarkov, D. and I. Horrocks (2006). "FaCT++ Description Logic Reasoner: System Description." *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006),* Vol. 4130: 292–297.

W3C. (2010a). "Semantic Web Activity, http://www.w3.org/2001/sw." Retrieved 15th August, 2010.

W3C (2010b). "W3C, Resource Description Framework (RDF), http://www.w3.org/RDF/ , last accessed September 10th 2010."

W3C. (2012). "OWL Web Ontology Language." Retrieved 12th June, 2012, *from http://www.w3.org/TR/owl-features/*.

WISDOM. (2011). "Water-Related Information System for the Sustainable development of the Mekong Delta." Retrieved 25, 2011, *from http://wisdom.eoc.dlr.de/en/content/objectives-wisdom-project*.

WISDOM. (2012). "WISDOM Information System." Retrieved 17th July, 2012, *from http://wisdom.eoc.dlr.de/Elvis/*.