ONTOLOGY DEVELOPMENT FOR INTEROPERABILITY OF OBSERVATION DATA

Masahiko Nagai, Masafumi Ono, Ryosuke Shibasaki

Earth Observation Data Integration and Fusion Research Initiative
The University of Tokyo, Japan,
435 Research Centers, CSIS, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan
Tel: +81-4-7136-4307 Fax: +81-4-7136-4292;
- nagaim@iis.u-tokyo.ac.jp

WG II/6 – System Integration and Interoperability

KEY WORDS: Database, Terminology, Data Mining, Global-Environmental-Databases, Interoperability

ABSTRACT:

The Ontology registry system is developed to collect, manage and compare ontological information for integrating global observation data. Data sharing and data service such as support of metadata deign, structuring of data contents, support of text mining are applied for better use of data as data interoperability. Semantic network dictionary and gazetteers are constructed as a trans-disciplinary dictionary. Ontological information are added to the system by digitalizing text based dictionaries, developing "knowledge writing tool" for experts, and extracting semantic relations from authoritative documents with natural language processing technique. The system is developed to collect lexicographic ontology and geographic ontology.

1. INGRODUCTION

The global observation is lying on trans-disciplinary fields, such as meteorology, hydrology, geology, geography, agriculture, biology, and so on. This is one of the most difficult factors to share the global observation data, because data user may need to understand several scientific fields. Under this trans-disciplinary condition, not only standardization of data structure but also communizing particular terminology and classification schema are serious hindrances to data sharing and integration of distributive data. If all systems use a standard model, various kinds of information can be integrated easily. However, arriving at a single standard requires enormous times and labors. That is, it is unrealistic to assume all models are standardized. Especially in earth observation data, distributive system is expected to utilize flexibly and easily with various needs. In this study, improving of the interoperability among the data is conducted in distributed or dispersed in space and different disciplines.

"Ontology" is originally used as Philosophical word, which means the branch of metaphysics that deals with the nature of being. But currently, in the field of context of knowledge sharing, the term ontology means a specification of a conceptualization. That is, ontology is a description of the concepts and relationships that can exist for a community or a particular field. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general (Smith, 2003).

In order to integrate or share global observation data, ontology registry system is developed to collect, manage, and compare ontological information such as data dictionaries, classification schemas, terminologies, thesauruses, and their relations. Data sharing and data service through the supporting of data retrieval, metadata deign, and text mining are applied for better or effective use of data. Semantic network dictionary is proposed as a trans-disciplinary dictionary. Dictionaries and data models are added to the system, developing "knowledge writing tool" for experts, and extracting semantic relations from authoritative documents with natural language processing technique. Generally, ontology is applied to strict and well defined implication such as task ontology (Kitamura, et al., 2004), but in this study, ontology is not restricted and applied as reference information for interoperability. Ontological information is classified into two groups, lexicographic ontology and geographic ontology, as shown in Figure 1. There are numerous amounts and different types of data. An individual database has its own definition of data; such kind definition is described as schema, for example, land use data schema, climate data schema. Under individual different data schema, it has different data names. Referring lexicographic ontology, it may estimate or sometimes successfully establish association of data. As long as, definition of data itself is focused on data interoperability, lexicographic ontology is used. But, if it focus on the location of the observation site, it is necessary to have a dictionary for geographic names for establishing association of data, so such kinds of ontology called geographic ontology. Thus, at least it is necessary to have two different types of ontology.

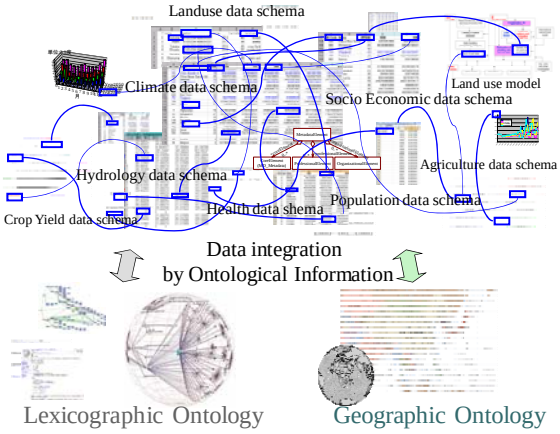

Figure 1. Ontological information

2. SEMANTIC NETWORK DICTIONARY

2.1 What is Semantic Network Dictionary?

Semantic network dictionary means that a certain term is expressed by definitions and relations of terms such as synonym, homonym. Entry words, definition, source, and author are handled as a node, and relations of terms are handled as a link. There are three peculiarities for the semantic network dictionary and its usage in terms of reliability, simple structure, and easy browsing and modification.

At first, the semantic network dictionary must be reliable, when users integrate data by referring information. If reliability is low, interoperability of data is not achieved. For reliability of the information, reliable data source should be applied, and data documentation must be obvious. In this study, collaboration with scientific society is conducted for data reliability. List of technical terms and association of terms are provided as ontological information from specialists. Reliability of data documentation is also achieved by adding authors and title of the references. Not only achieving technical terms but also editing of terms is carried out by specialist for data reliability.

Secondly, semantic network dictionary consists of technical terms and their relations, so the basic structure is quite simple. That is, it is easy to obtain a lot of data from various sources, and it helps to save labor for data construction. This is one of the key points to collect ontological information.

Thirdly, the purpose of semantic network dictionary is to support interoperability of data set, that is, it is necessary to refer to trans-disciplinary field easily. Structure of dictionary is just network between technical terms, so browsing is very simple like hyper link of web browser. Also, it is easy to add or edit their links and nodes, and to cut off certain parts of dictionary, and to dump in XML format.

2.2 Semantic MediaWiki

In order to collect the lexicographic ontology with above peculiarities, semantic network dictionary is developed based on Semantic MediaWiki, SMW (Leuf and Cunningham, 2001), which allows users to freely create and edit contents using any Web browser. SMW is a feature-rich wiki implementation, as shown in Figure 2. SMW handles hyperlinks and has simple text syntax for creating new pages and crosslink between terms. In SMW, a visual depiction of content is expressed by tags. It is not easy to add relations by tags. Therefore, in this study, table like editor is developed by modification of original SMW. SMW displays not only definition, but also relations of terms. Table editor is applied in order to modify relations of terms by using a table without putting tags. Also, the data managing system is developed with SMW, because it is necessary to maintain data reliability. The data managing system controls user access and gives permission to add or edit the semantic network dictionary.

In this study, dictionaries, and data schemas are collected for examination of semantic network dictionary by using SMW. The fields of collected trans-disciplinary dictionaries are agriculture, biology, civil engineering, earth science, soil science, meteorology, health science, and remote sensing. Moreover, landuse classification schemas are collected. These scientific fields are considered as major fields for global observation data and it is necessary to understand technical terms and their relations for data integration among the fields.

Figure 2. Semantic MediaWiki

2.3 Reverse Dictionary

Constructed ontological information is used for the reverse dictionary. Reverse dictionary describes a concept of term from definition and association of terms. Reverse dictionary is developed based on GETA which is developed National Institute of Informatics, Japan. It is tools for manipulating large dimensional sparse matrices for text retrieval. GETA is an engine for association's calculation such as similarity measurement (Takano, et al., 2000). For example, user wants to know about "instrument to indicate level of water". Reverse dictionary returns the list of terms with similarity scores, such as "Water-level recorder", "Inclined gauge; inclined gage", "Staff gauge; staff gage", and so on. Reverse dictionary relates data by calculation of similarity.

2.4 KeyGraph viewer

In order to compare associations among the different key words, graph representation is useful as shown in Figure 3. Landuse classification schema in Thailand and Indonesia is compared as an example. The term "water body" can be found in both countries. Apparently, both landuse classes look the same, but level of hierarchy is a bit different in each classification schema. In the case of Indonesian landuse, "water body" does not include water course, but "water body" in Thailand includes all water related land types. Consequently, graph representation proves a clear distinction between the two terms.
Now the new information such as relations of "water body" in both countries can be developed. These kinds of information are treated as new ontological information, and added to the ontology registry system through the Semantic Media Wiki. The ontological information can grow autonomously by adding relations, and then it will be more and more useful.

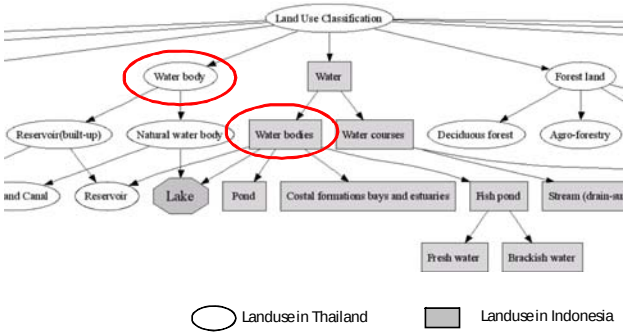

Figure 3. KeyGraph viewer

3. GAZETTEER

3.1 What is Gazetteer?

Gazetteer is developed for geographic ontology as a part of ontology registry system. Gazetteer is defined as an important reference for information about places and place names used in conjunction with an atlas (Hill, et al., 1999). In order to integrate global observation data, the system is constructed associations of place names. In this study, place names with latitude and longitude are collected as ontological information. For ontological information, it is necessary to collect truthful contents, so not only data construction, but also system management are considered as a part of the system.

The basic of gazetteer is the correspondence between place names and spatial information. Place name is usually used in linguistic activity of human. In the society, it is sophisticated information to exchange or distribute information. Figure 4 shows the concept of gazetteer which is developed in this study.

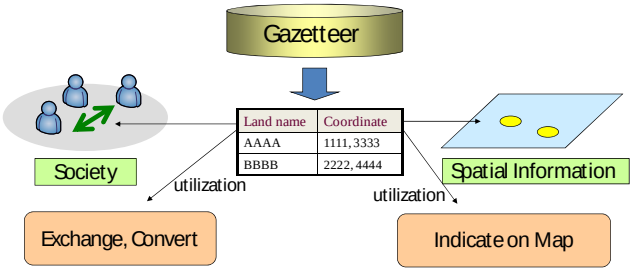

Figure 4. Concept of Gazetteer

	Items	Descriptions
Primitive	Place name	Place name in particular language.
	Coordinate (point)	Latitude and longitude in particular coordinate system.
Mandatory	Language	Language for data base. e.g. English, Chinese, Japanese, etc.
	Country	Country name which exists place name.
	Category	Place name category. e.g. city name, political boundary, mountain, river, etc.
	Life cycle	Life cycle for using place name. e.g. 1990-04-01～2007-03-31.
	Editing histories	Editing record for editor. e.g. register on 2007-04-01, etc.
Optional	Second names	Other names for place name.
	English name	English expression if place name is not written in English.
	Scale	Scale to show on the Map.
	Relations	Relation of other place names. e.g. place name A contains place name B, place name A is near to place name C, etc.
	MBR (Minimum Bounding Rectangle)	The broad expanse of place name in Latitude and Longitude.
	Image files	Related image file. e.g. landscape photos, maps, etc.
	Notes	Free text.

Table 1. Items for gazetteer

In this study, the system to browse and modify place name data by using GUI is developed. In order to collect as high quality data, user management system is also developed. The gazetteer system can be used globally without limitation of spatial scale, region, and language. As an I/F function, Google Maps is used, as shown in Figure 5. Input data can be listed on Google Maps, and the data can be retrieved from the map, and also coordinate information can be modified by using the map.

One of the key points of collecting the information is determination of items for database as attributes which are related to place name and map. Table 1 shows the list of items for gazetteer in this study. The items classified into 3 groups, primitive items, mandatory items, and optional items. Those items are ontology regarding to geographic information.

3.2 Gazetteer system design

As an initial data, GNS data (GEOnet Names Server) is applied. GNS provides access to NGA (National Geospatial-Intelligence Agency's) and the BGN (U.S. Board on Geographic Names') database of geographic feature names and locations for locations of all over the world. The data is the official repository of place name, approximately 8,000,000 points of data. The place name data can be found by structure like grid cells, which means that resolution and accuracy of the coordinate is limited. In that sense, it may be necessary to improve such kinds of initial data by the system.

In order to operate information, the gazetteer system has three types of functions for users in terms of a visitor, an editor, and a manager, and GUI for each users are developed and associated with Google Maps API. A visitor retrieves from the registered data and refers the information. An editor can registers, modifies, and deletes the data. Only authorized users can edit in order to maintain reliably as same as Semantic Media Wiki. A manager administers the data set and users on the system.

In the system, there are two types of retrieval, one is item retrieval, and another is map retrieval. The retrieval from items is conducted by using place name, latitude and longitude, and so on. Retrieval from map is conducted by using boundary rectangle. All the place names in the rectangle box are picked up as shown in Figure 5.

At this function, a new place name is added, if there is no information of the object. All the items should be added together with place name and coordinate information. Coordinate information can be acquired from Google Maps. If the information is not good enough to express geographic ontology, the data can be modified. Coordinate information and scale can be modified from the map. MBR for right upper and left lower coordinate are set by using MBR bar.

Information of data and user is administrated by system managers. Their roles are input data and user collectively in order to assemble reliable information. They control and certificate an account for users. Also, the data is imported and exported for this system. One important function is checking of the log for other function. Not only watching of prohibited process but also retrieved terms are recorded. If users try to retrieve a data but no information about place name, a manager can recognize non-registered place name. Then, that place name is added for users.

Constructed geographic ontology is used for the reference information for interoperability. For integration of observation data, it is essential to clarify topological information, such as contain, near, part-of, is-on, consist-of, and so on. The gazetteer collects that topological information by association with Google Maps.

Figure 5. Land name retrieval

4. ONTOLOGY REGISTRY

4.1 What is ontology registry?

Ontology registry is comprehensive, authoritative reference for information about data model, data specification, data definition, and their relation of observation data, as shown in Figure 6. The ontology registry supports the creation and implementation of data model that are designed to encourage the efficient sharing of observation data. The ontology registry catalogues for data elements in application systems. The ontology registry does not contain observation data, but it may show data availability. It provides descriptive information to make the data model. The ontology registry links with the semantic network dictionary and gazetteer. In order to describe data element, it is necessary to understand technical terms and geographic information. The semantic network dictionary and gazetteer is considered as part of the ontology registry system.

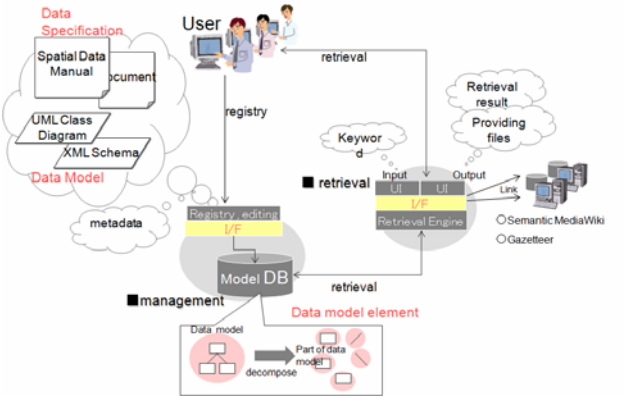

Figure 6. Concept of ontology registry

4.2 Use case of ontology registry system

Figure 7 shows the use case of ontology registry. Data model, data specification, data definition, and their relation of observation data is collected and managed in ontology registry system. When user creates data model X, the user can refer existing data model A, not only data item, but also data inheritance and relation. This is very important because if user create new data model without referring existing model, many different sort of models are jumbled up. Also, if ontology registry system archive data models with their relations and definitions, it is very helpful to integrated actual data.

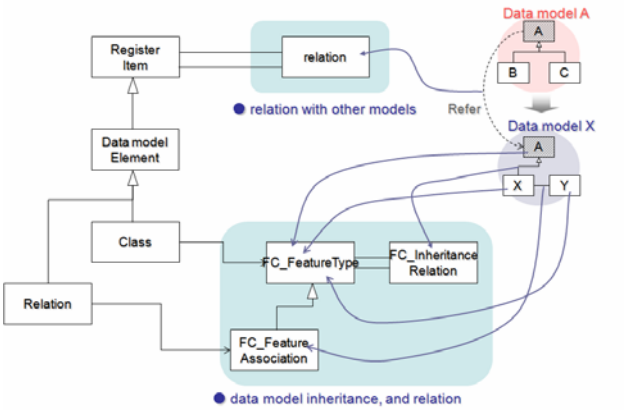

Figure 7. Use case

5. CONCLUSIONS

In conclusion, many standardization organizations are working for syntactic level of the interoperability, but in the same time, semantic interoperability of data must be considered in heterogeneous condition and also very diversified and large volume of data set for global observation data. Ontological information is developed by the proposed system as lexicographic ontology and geographic ontology. This is very challenging method with collaboration from scientists in different background and language; therefore, it is very important to develop effective tools to collect reliable ontological information.

Ontology registry system with semantic network dictionary and gazetteer are developed to register and update of ontological

information based on Semantic MediaWiki and Google Maps. They must be a tool to support scientist and specialist for their ontology development. In order to invite contributions from the user community in various scientific fields, it is necessary to provide more sophisticated and user friendly tools and systems for sustainable development of ontological information.

REFERENCES

Hill, L., Frew, J., Zheng, Q., 1999, Geographic names: The implementation of a gazetteer in a georeferenced digital library. D-Lib, January.

Kitamura, Y,. Kashiwase, M., Fuse M., and Mizoguchi, R., 2004, Deployment of an ontological framework of functional design knowledge, Advanced Engineering Informatics, Volume 18, Issue 2, pp. 115-127.

Smith, B., 2003, Preprint version of chapter “Ontology”, in L. Floridi (ed.), Blackwell Guide to the Philosophy of Computing and Information, Oxford: Blackwell, pp.155–166.

Takano, A., Niwa, Y., Nishioka, S., Iwayama, M., Hisamitsu, T., Imaichi, O., Sakurai, H., 2000, Information Access based on Associative Calculation, In Lecture Notes in Computer Science LNCS:1963, Springer.

Leuf, B. and Cunningham, W., 2001, The Wiki Way: Quick Collaboration on the Web. Addison-Wesley, USA.