TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 54/67 pp336-342
Volume 13, Number S1, October 2008

Interoperability for Global Observation Data by Ontological Information*

Masahiko Nagai**, Masafumi Ono, Ryosuke Shibasaki†

Earth Observation Data Integration and Fusion Research Initiative (EDITORIA),
The University of Tokyo, Chiba 277-8568, Japan;
†Center for Spatial Information Science, The University of Tokyo, Tokyo 153-8904, Japan

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 is 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.

Key words: ontology, interoperability, data integration, gazetteer, semantic network dictionary

Introduction

The global environment is lying on trans-disciplinary fields such as meteorology, hydrology, geology, geography, agriculture, and biology. Under the 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 recently, 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[1].

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, and thesauruses. 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. 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

Received: 2008-05-31

* Supported by the Data Integration and Analysis System (DIAS) Project

** To whom correspondence should be addressed.
E-mail: nagaim@iis.u-tokyo.ac.jp; Tel: +81-4-7136-4307

from authoritative documents with natural language processing technique. Generally, ontology is applied to strict and well defined implication such as task ontology[2], but in this study, ontology is applied as reference information for interoperability. Ontological information is classified into two groups, lexicographic ontology and geographic ontology, as shown in Fig. 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 and 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.

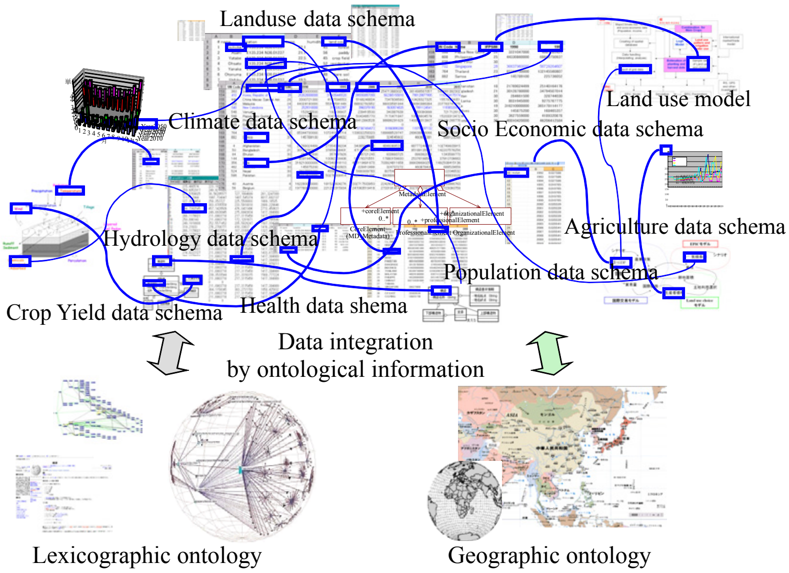


Fig. 1 Two ontological information

1 Ontology Development

In order to develop ontological information, there are four stages to process, in terms of registration, browsing, modification, and utilization. The ontological information collects, edits, provides, and supports the uses through the system.

1.1 Registry

In this stage, ontological information is collected and registered to the system. In order to establish initial stock of knowledge, it is necessary to tackle with piles of papers. If data is available in digital, it is not difficult to register. However, most of original data is available in paper. In this case, original paper based data are scanned and processed by OCR. This work is labor demanding process. Individual keyword, pronunciation, author, reference word, and page number must be structured using XML. Pre-processed data are converted into XML format and registered.

Also, gazetteer which is the list of geographic names is collected with longitude and latitude coordinate. For example in this study, gazetteer is originally developed by National Geospatial-Intelligence Agency (NGA) and is available in digital[3].

1.2 Browsing

Browsing tools are developed for clarifying ontological expression and visualizing relationship between terms. It provides reference information to users visually. The graph representation is conducting for showing association among the different key words. The related terms link such as is-a, part-of, synonym, and antonym. The graph representation is also used for browsing ontological information with meanings and associations with other words and concepts.

1.3 Modification

Registered information is modified or edited in this system. The registry system is directly linked to Semantic Media Wiki. Semantic Media Wiki allows various experts to edit ontological information. Table editor is developed for simple modification. Table editor converts table data to XML format automatically.

Also, Google Maps are very convenient for gazetteer to add or edit points with geographic attributes. Gazetteer database links to Google Maps by using Google Maps API. Relationships or descriptions of the terms are added by using these softwares or Web services.

1.4 Utilization

In this stage, ontological information is provided to support the uses of ontology. Ontological information is utilized for data integration, information retrieval, reconstruction of data model, and so on. Reverse dictionary is developed as one of the sample usage of ontological information. Reverse dictionary is a dictionary which gets back a list of index words from list of key words or phrases related to that concept. As the result of user analysis, new knowledge may be added. Additional knowledge of ontological information is registered here again. Then, newly created ontology is returned to the first stage.

2 Semantic Network Dictionary

2.1 What is semantic network dictionary?

Semantic network dictionary means that a certain term is expressed by definition 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.

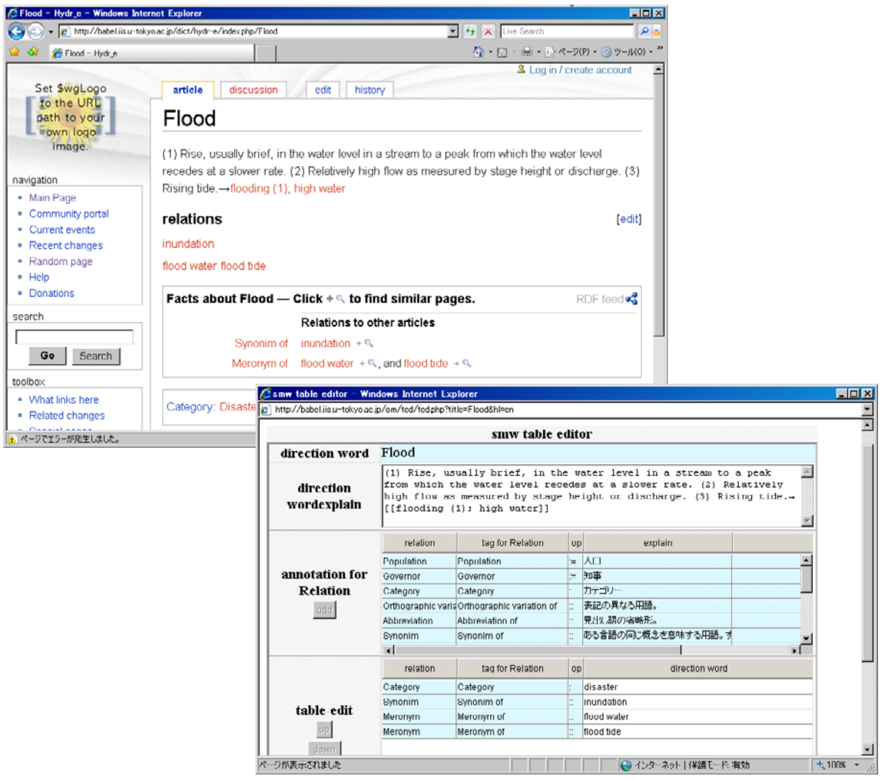
At first, the ontological information must be reliable, when users integrate data by referring semantic network dictionary. If reliability is low, interoperability of data is not achieved. For reliability of the information, reliable data source should be applied, and data documentation must be obvious. 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 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 terms, so browsing is very simple like hyper link of web browser. Also, it is easy to add or edit their links and nodes, to cut off certain part of dictionary, and to dump in XML format.

2.2 MediaWiki

In order to collect the lexicographic ontology with above peculiarities, semantic network dictionary is developed based on MediaWiki[4], which allows users to freely create and edit contents using any Web browser. MediaWiki is a feature-rich wiki implementation. MediaWiki handles hyperlinks and has simple text syntax for creating new pages and crosslink between terms. In MediaWiki, 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 MdiaWiki. Figure 2 shows MediaWiki and table editor, which brows explanation of the term. MediaWiki 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. In this study, dictionaries and data schemas are collected for examination of semantic network dictionary by using MediaWiki. 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.



Flood – Hydr_e – Windows Internet Explorer

http://babel.iis.u-tokyo.ac.jp/dict/hydr-e/index.php/Flood

Flood

(1) Rise, usually brief, in the water level in a stream to a peak from which the water level recedes at a slower rate. (2) Relatively high flow as measured by stage height or discharge. (3) Rising tide.→flooding (1); high water

relations

inundation

flood water flood tide

Facts about Flood — Click + to find similar pages.

Relations to other articles

Synonim of	inundation +
Meronym of	flood water +, and flood tide +

Category: Disaste

ページでエラーが発生しました。

smw table editor – Windows Internet Explorer

http://babel.iis.u-tokyo.ac.jp/om/tod/tod.php?title=Flood&hl=en

smw table editor

direction word	Flood
direction wordexplain	(1) Rise, usually brief, in the water level in a stream to a peak from which the water level recedes at a slower rate. (2) Relatively high flow as measured by stage height or discharge. (3) Rising tide.→ [[flooding (1); high water]]

annotation for Relation

relation	tag for Relation	op	explain
Population	Population	:=	人口
Governor	Governor	:=	知事
Category	Category	:	カテゴリー
Orthographic varia	Orthographic variation of	::	表記の異なる用語。
Abbreviation	Abbreviation of	::	見出し語の省略形。
Synonim	Synonim of	::	ある言語の同じ概念を意味する用語。す

table edit

relation	tag for Relation	op	direction word
Category	Category	:	disaster
Synonim	Synonim of	::	inundation
Meronym	Meronym of	::	flood water
Meronym	Meronym of	::	flood tide

ページが表示されました

Fig. 2 MediaWiki based system

2.3 KeyGraph viewer

In order to compare associations among the different key words, graph representation is useful as shown in Fig. 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 are 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 useful.

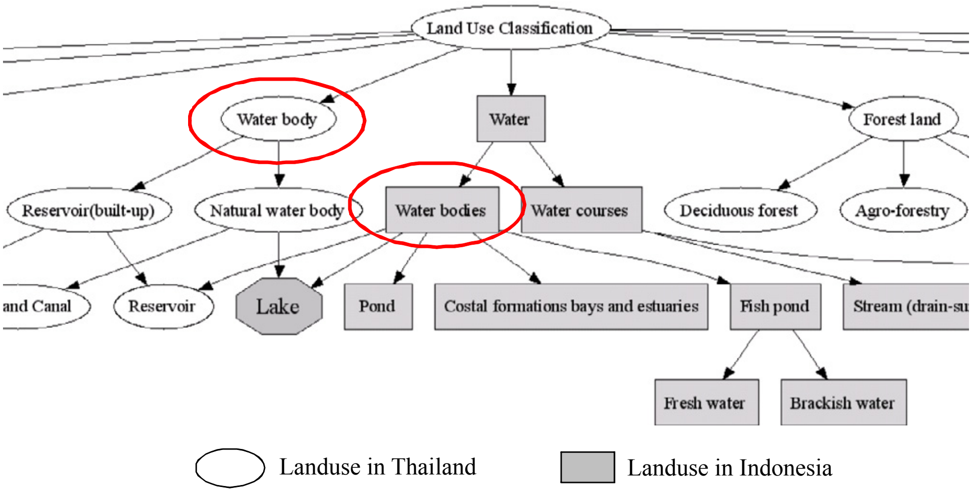
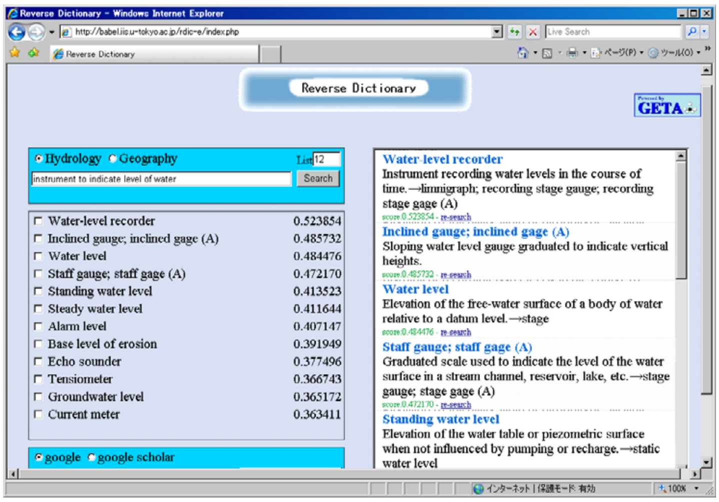


Fig. 3 KeyGraph viewer

2.4 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[5,6].

Figure 4 shows reverse dictionary. 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.


Reverse Dictionary - Windows Internet Explorer
http://babel.iis.u-tokyo.ac.jp/rdic-e/index.php
Live Search
Reverse Dictionary
ページ(P)
ツール(O)
Reverse Dictionary
Powered by GETA
インターネット | 保護モード: 有効
100%

● Hydrology ○ Geography List 12

instrument to indicate level of water [Search]

□ Water-level recorder	0.523854
□ Inclined gauge; inclined gage (A)	0.485732
□ Water level	0.484476
□ Staff gauge; staff gage (A)	0.472170
□ Standing water level	0.413523
□ Steady water level	0.411644
□ Alarm level	0.407147
□ Base level of erosion	0.391949
□ Echo sounder	0.377496
□ Tensiometer	0.366743
□ Groundwater level	0.365172
□ Current meter	0.363411

● google ○ google scholar

Water-level recorder
Instrument recording water levels in the course of time.→limnigraph; recording stage gauge; recording stage gage (A)
score 0.523854 - re-search

Inclined gauge; inclined gage (A)
Sloping water level gauge graduated to indicate vertical heights.
score 0.485732 - re-search

Water level
Elevation of the free-water surface of a body of water relative to a datum level.→stage
score 0.484476 - re-search

Staff gauge; staff gage (A)
Graduated scale used to indicate the level of the water surface in a stream channel, reservoir, lake, etc.→stage gauge; stage gage (A)
score 0.472170 - re-search

Standing water level
Elevation of the water table or piezometric surface when not influenced by pumping or recharge.→static water level

Fig. 4 Reverse dictionary

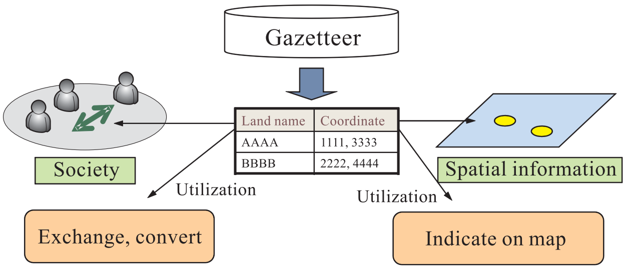
3 Gazetteer

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[7]. In order to integrate global observation data, the system is constructed by 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 is 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 5 shows the concept of gazetteer which is developed in this study.

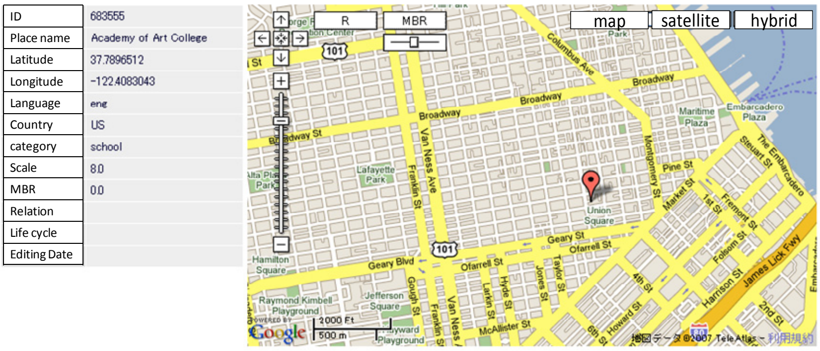
In this study, the system to browse and modify place name data by using GUI is developed. In order to collect high quality data, user management system is also developed. The gazetteer system can be used globally without limitation of spatial scale, region, and language. As an I/F function, Google Maps is used as shown in Fig. 6. 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.



Land name	Coordinate
AAAA	1111, 3333
BBBB	2222, 4444

Fig. 5 Concept of gazetteer

One of the key points of collecting the information is determination of items for database as attributes which are related to palace name and the 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.



ID	683555
Place name	Academy of Art College
Latitude	37.7896512
Longitude	-122.4083043
Language	eng
Country	US
category	school
Scale	8.0
MBR	0.0
Relation	
Life cycle	
Editing Date	

Fig. 6 Gazetteer

Table 1 List of items for 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, and Japanese.
	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.
	Minimum Bounding Rectangle (MBR)	The broad expanse of place name in Latitude and Longitude.
	Image files	Related image file. e.g., landscape photos, maps, etc.
	Notes	Free text.

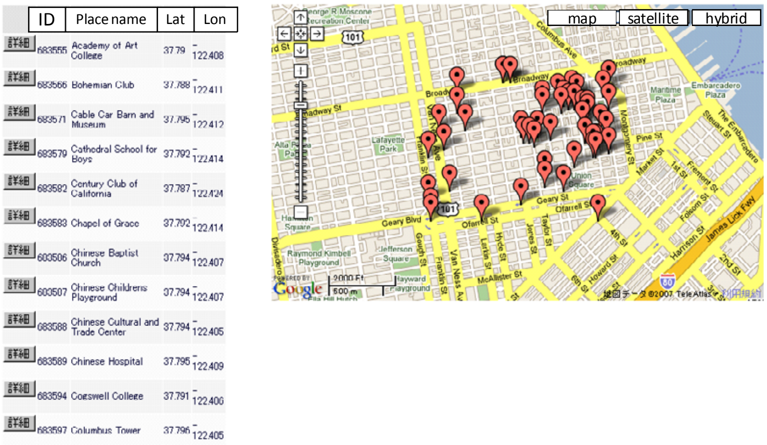
3.2 Gazetteer system

As an initial data, GNS data (GEOnet Names Server; http://earth-info.nga.mil/gns/html/index.html) is applied. GNS provides access to National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names' (BGN) 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 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 the other 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 Fig. 7.

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.



	ID	Place name	Lat	Lon
詳細	683555	Academy of Art College	37.79	-122.408
詳細	683566	Bohemian Club	37.788	-122.411
詳細	683571	Cable Car Barn and Museum	37.795	-122.412
詳細	683579	Cathedral School for Boys	37.792	-122.414
詳細	683582	Century Club of California	37.787	-122.424
詳細	683583	Chapel of Grace	37.792	-122.414
詳細	600506	Chinese Baptist Church	37.794	-122.407
詳細	600507	Chinese Childrens Playground	37.794	-122.407
詳細	683588	Chinese Cultural and Trade Center	37.794	-122.405
詳細	683589	Chinese Hospital	37.795	-122.409
詳細	683594	Cogswell College	37.791	-122.406
詳細	683597	Columbus Tower	37.796	-122.405

Fig. 7　Land name retrieval

Information of data and user is administrated by system managers. Their roles are inputting 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 retrieving of 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 visitor.

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. GEND collects that topological information by association with Google Maps.

4　Conclusions and Future Work

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. Ontological information is developed by the proposed system as lexicographic ontology and geographic ontology. This is a very challenging method with collaboration from scientists in different background and language; therefore, it is very important to develop effective tools to develop ontological information.

Semantic network dictionary and gazetteer are developed to register and update of ontological information based on MediaWiki and Google Maps. They are tools 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.

Acknowledgments

This study is supported by Data Integration and Analysis System (DIAS) project. DIAS project is designated as key technology of national importance, Japan, which contributes for GEOSS.

References

[1] Barry S. Preprint version of chapter “Ontology”. In: Floridi L ed. Blackwell Guide to the Philosophy of Computing and Information. Oxford: Blackwell, 2003: 155-166.

[2] Yoshinobu K, Masakazu K, Masayoshi F, et al. Deployment of an ontological framework of functional design knowledge. *Advanced Engineering Informatics*, 2004, **18**: 115-127.

[3] Peter L, Magesh B, Dominic W, et al. The information commons gazetteer—A public resource of populated places and worldwide administrative divisions. In: Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy, 2006: 24-26.

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

[5] Akihiko T, Yoshiki N, Shingo N, et al. Information Access based on Associative Calculation. In: Lecture Notes in Computer Science LNCS: 1963. Springer, 2000.

[6] An associative search system based on a Generic Engine for Transposable Association (GETA), IPSJ SIG Notes, Vol.2000, No.53(20000601), 2000.

[7] Hill L L, Frew J, Zheng Q. Geographic names: The implementation of a gazetteer in a georeferenced. D-Lib, 1999.