EARTH OBSERVATION DATA INTEROPERABILITY ARRANGEMENT WITH ONTOLOGY REGISTRY

M. Nagai [a, *], M. Ono [b], R. Shibasaki [b]

[a] Geoinformatics Center, Asian Institute of Technology, Km.42, Paholyothin Highway, Klong Luang, Pathumthani, 12120, Thailand – nagam@ait.asia
[b] Center for Spatial Information Science, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan - maono@iis.u-tokyo.ac.jp, shiba@csis.u-tokyo.ac.jp

Commission ICWG II/IV

KEY WORDS: Semantic Interoperability and Ontology for Geospatial Information

ABSTRACT:

Standardization organizations are working for syntactic and schematic level of interoperability. At the same time, semantic interoperability must be considered as a heterogeneous condition and also very diversified with a large-volume data. The ontology registry has been developed and ontological information such as technical vocabularies for earth observation has been collected for data interoperability arrangement. This is a very challenging method for earth observation data interoperability because collaboration or cooperation with scientists of different disciplines is essential for common understanding. Multiple semantic MediaWikis are applied to register and update technical vocabularies as a part of the ontology registry, which promises to be a useful tool for users. In order to invite contributions from the user community, it is necessary to provide sophisticated and easy-to-use tools and systems, such as table-like editor, reverse dictionary, and graph representation for sustainable development and usage of ontological information. Registered ontologies supply the reference information required for earth observation data retrieval. We proposed data/metadata search with ontology such as technical vocabularies and visualization of relations among dataset to very large scale and various earth observation data.

1. INTRODUCTION

The global environment is lying on trans-disciplinary fields, such as meteorology, hydrology, geology, geography, agriculture, biology, and so on. It is essential to cross these trans-disciplinary fields for measuring the global environmental problems, such as climate change, global worming, various disasters, pollutions and so on. One of the key issues is data interoperability arrangement under the trans-disciplinary condition. There are three aspects of the data interoperability: syntactic interoperability, schematic interoperability and semantic interoperability. Improvement of those aspects of interoperability is needed for integrated use of heterogeneous data. To improve the syntax interoperability, many efforts have already been made such as standardization of data formats and development of XML-based data encoding rules, for example, ISO (International Organization for Standardization) standard and OGC (Open Geospatial Consortium) standard. Improvement of schematic interoperability is to make accessible geological map data in a common data format. Improvement of semantic interoperability requires common understanding among different ontologies, terminologies, taxonomies, including definitions and associations of various vocabularies, concepts/terms, name spaces, classification schemes and so on, which is collectively called an "ontology". The word "ontology" was originally used in philosophy, to refer to the branch of metaphysics that deals with the nature of being. Currently, in context of knowledge sharing, the term means a specification of a conceptualization (Gruber, 1993). In recent years, several institutions have initiated efforts to propose a standard ontology and/or terminology/taxonomy related with Earth Observation. SWEET (Semantic Web for Environment and Technology) by NASA (National Aeronautics and Space Administration) is one of such ontology (NASA, 2011). FAO (Food and Agriculture Organization of the United Nations) is making similar kinds of efforts based on AGROVOC, that is s a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (FAO, 2011). Many other ontologies and terminologies/taxonomies are expected to be proposed by other expert/professional communities and institutions. For semantic data interoperability, ontological information including terminology, taxonomy, glossary, etc., must be collected, managed, referred and compared; for example, data dictionaries, classification schemata, terminologies, thesauruses, and their relations are handled. Common understanding of heterogeneous semantic information is used for data sharing and data services such as supporting data retrieval, metadata design, information mining, and so on.

In this study, the ontology registry is constructed for information sharing by using a Semantic MediaWiki, which helps to gather ontological information, terminologies, scientific vocabularies and their associations for data interoperability among diversified and distributed data sources. Generally, ontology is applied to a strict and well-defined purpose, classes and instances such as a task ontology (Kitamura, et al., 2004), but in this study, the scope of ontologies is not restricted and comprises any reference information based on terminology of technical vocabulariess for data interoperability. The ontology registry creates a "knowledge writing tool" for experts, by extracting semantic relations from authoritative documents using natural language processing techniques, such as

* Corresponding author. This is useful to know for communication with the appropriate person in cases with more than one author.

morphological analysis and semantic analysis for earth observation data interoperability.

2. EXISTING REGISTRY

There are several registries as earlier studies. First of all, existing data registries were reviewed to clarify necessary requirements. GBIF (Global Biodiversity Information Facility) is an information portal for global biodiversity data that is initiated by OECE (Organization for Economic Co-operation and Development). Diversified and distributed biodiversity data can be retrieved through clearing house system with common indexes. There is no concept to handle multi-disciplinary field and to register several different data specifications, but GBIF provides the tool to convert from local format to common standard format of the registry (GBIF, 2011).

EPA SoR (Environmental Protection Agency System of registries) provides a gateway and search capability to several registries and repositories. These registries include a link in EPA's information architecture to exchange the Agency's data efficiently with Agency's standards program. This registry handles legacy data and mainly contains academic information and observation data (EPA, 2011).

METeOR (Australian Institute of Health and Welfare Metadata Online Registry) is a repository for Australian national metadata standards for health, housing and community services statistics and information with ISO11179. This is knowledge-based registry to be possible retrieval of legacy information (Australian Institute of Health and Welfare, 2011).

UK ITS (UK Highway Agency ITS Metadata Registry) is very typical registry with data definitions and data models, with an associated supporting process for improvement of quality and for harmonization across different systems. The registry aims to cut across work among different system and avoid duplication for data interoperability (UK Highways Agency, 2011).

All registries are possible to retrieve by keywords, but efficiency of retrieval is not good enough if scale of registries is huge. Therefore, data retrieval is supported though advanced and hierarchical search, alphabetical order. In case of GBIF, scientific name of the portal index can be linked with common name for supporting data retrieval. In case of METeOR, keywords are selected in accordance with ISO11179. In addition, those registries are domain specific registry; therefore they support data description and retrieval for interoperation in a specific domain. In this paper, we propose earth observation data arrangement with flexible semantic description. It uses a simple inheritance mechanism to provide multiple ontologies for a heterogeneous condition and also very diversified with a large-volume data.

3. DESIGN FOR ONTOLOGY REGISTRY

The ontology registry has been developed to store ontological information that a certain term is expressed by definition and relations to other terms such as is-a, part-of, synonym, homonym, and so on. Entry words, definitions, sources, and authors are handled as nodes, and relations to other terms are handled as links. Those terms are surrounded by other relational terms. There are a few key requirements of the ontology registry; reliability, simple structure, and easy browsing and modification.

3.1 Reliability

The semantic information in the ontology registry must be reliable, when users retrieve data by referring to ontological information. If reliability is not confirmed, the interoperability is not achieved. For reliability of the information, reliable data sources should be selected, and data documentation must be obvious. In this study, collaboration with scientific society and international organization is made for reliability. Lists of technical vocabularies and associations of each term are provided as ontological information from specialists of each field. Reliability of data documentation is also achieved by adding authors and titles of the references. Editing, as well as definition, of terms is followed by original sources for keeping reliability. The initial sources of the registry are listed below.

GCMD (NASA's Global Change Master Directory) science keyword is a hierarchical set of Earth science keywords. Each metadata has at least one or more keywords that a user can find datasets. GCMD keywords have been classified into following categories; agriculture, atmosphere, biosphere, human dimensions, hydrosphere, land surface, oceans, climate, radiance/imagery, snow or ice, solid earth, and sun-earth interactions (NASA Goddard Space Flight Center, 2011).

CF Standard names includes a precise definition of each quantity named and has more than 1900 entries. The standard name is also independent of units of measure and coordinate variables used to describe its variation (NetCDF Climate and Forecast Metadata Convention, 2011).

CEOS (the Committee on Earth Observation Satellites) database has two contents of CEOS database of missions, instruments and measurement (MIM) database and CEOS systems engineering office (SEO) database. MIM database is maintained by ESA (European Space Agency) and provides information for supporting the future earth observation mission, instrument and measurements plans. SEO database is designed to support CEOS strategic planning and gap assessments (CEOS Missions, 2011).

WMO Space Program glossary is a glossary provided by World Meteorological Organization. This glossary contains abbreviations and official names of satellites and sensors (WMO, 2011).

GEMET (GEneral Multilingual Environmental Thesaurus) is a multilingual thesaurus that can translate to 28 different languages. This environmental thesaurus is managed by the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA, 2011).

SWEET is an ontology managed by NASA and provides the basis for Earth system science phenomena and realms. SWEET is written in OWL format.

CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science) is an ontology in hydrology domain developed by a US academic community. It provides concepts and vocabularies for using hydrologic data (CUAHSI, 2011).

JAXA satellite glossary is a Japanese satellite related glossary provided by JAXA (Japan Aerospace Exploration Agency). This is published in JAXA web pages (JAXA, 2011).

Remote sensing glossary is published by Japan Association of Remote Sensing as a hard copy handbook. Therefore, this hard copy was digitized manually and created the digital version (Japan Association of Remote Sensing, 1989).

GIS glossary is developed by GIS Association of Japan. Its terms are mainly written in Japanese but some are written in English (GIS Association of Japan, 2011).

3.2 Simple Structure

The ontology registry consists of terms with definitions and their relations, so the basic structure of ontological information must be quite simple. This makes it 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 in collecting and managing ontological information. Original formats of ontological information are not only text and spreadsheet table but also XML (Extensible Markup Language), RDF (Resource Description Framework), and OWL (Web Ontology Language). These formats are simply converted and expressed for semantic uses.

3.3 Easy Browsing and Modification

The purpose of the ontology registry is to support the interoperability of data by making it easy to refer a trans-disciplinary field. The structure of such a registry is a simple network among vocabularies; so browsing the dictionary resembles operating a hyper link of web browser and web API (Application Program Interface). Also, it is easy to add or edit their links and nodes, and to cut off certain part of dictionary, and to export in XML and RDF format.

4. IMPLEMEMTATION FOR ONTOLOGY REGISTRY

In order to collect ontological information with above requirements, a registration system is developed based on Semantic MediaWiki (version SMW1.2). Semantic MediaWiki is a feature-rich wiki implementation. Semantic MediaWiki handles hyperlinks and has simple text syntax for creating new pages and cross-links between vocabularies (http://semantic-mediawiki.org/wiki/Semantic_MediaWiki). Entry words, definitions, sources, and authors are handled as nodes with tags, and relations to other terms are handled as links. Those terms are surrounded by other relational terms. Here, the separate Wiki manages each ontology or terminology, for example, SWEET Wiki is created for SWEET ontology, and GEMET Wiki is created for GEMET terminology.

4.1 Registration

At first, ontological information is registered to Semantic MediaWiki by the developed tool automatically by converting form text, RDF, and OWL to XML and importing to Wiki. Sometimes, ontological information is manually registered from book and Web pages. These existing dictionaries or glossary are already considered as ontological information. OCR (Optical character reader) is sometimes used to digitize the old sources. Secondly, symbols and abbreviations, such as related words and synonyms are extracted from the dictionary and converted from semantic structure to syntactic structure by natural language processing. Finally, imported ontological information is modified and revised by users with editing function as shown in Figure 1.

In Semantic MediaWiki, a tag is a visual depiction for text based contents. It is not easy to add or select appropriate relations by tags without knowledge of computer science and misspelling, so in this study, the table like editor was developed as a wiki plug-in for editing. The table editor links to editing page of the Wiki by pop-up window and suggest appropriate tags to control community authoring. The table editor is implemented by dhtmlGrid, v1.2 standard. XML is prepared for Web server. Semantic Media Wiki displays not only definitions, but also relations of terms among multiple Wikis. The table editor is applied in order to modify relations of terms by using a table without tags.

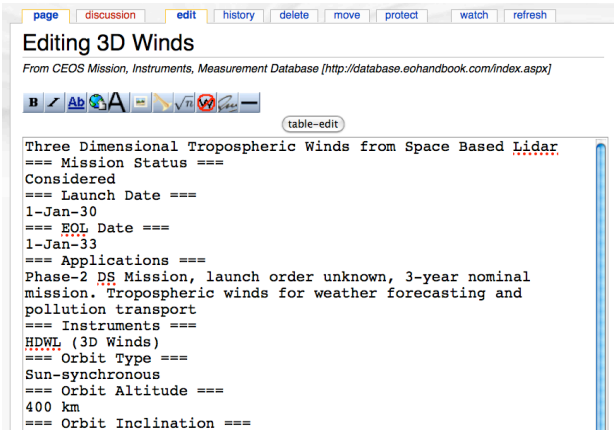

Figure 1. Semantic MediaWiki editing interface

4.2 Information Retrieval

Registered ontological information in Semantic MediaWikis is retrieved by the reverse dictionary. The reverse dictionary describes a concept of a term from definitions and associations of terms. The reverse dictionary is developed based on GETA (Generic Engine for Transposable Association), which was developed by the National Institute of Informatics, Japan (Takano, et al., 2000). It comprises tools for manipulating large-dimensional sparse matrices for text retrieval through more than one Wiki in all together. GETA is an engine for the calculation of associations such as similarity measurement of multiple Wikis. In order to create matrices to find similarity, morphological analysis is conducted for word segmentation and listing of ignored words for calculation. In the reverse dictionary, GETA can be directly used to realize associative searching systems, which accept a group of texts as queries, and return highly related texts in the relevance order. The usefulness of this retrieval is to cover the weakness of ordinary key-word search that sometimes returns no hit and sometimes returns too many hits.

As an example of information retrieval, suppose a user wants to know about a “satellite for sea surface temperature”. The

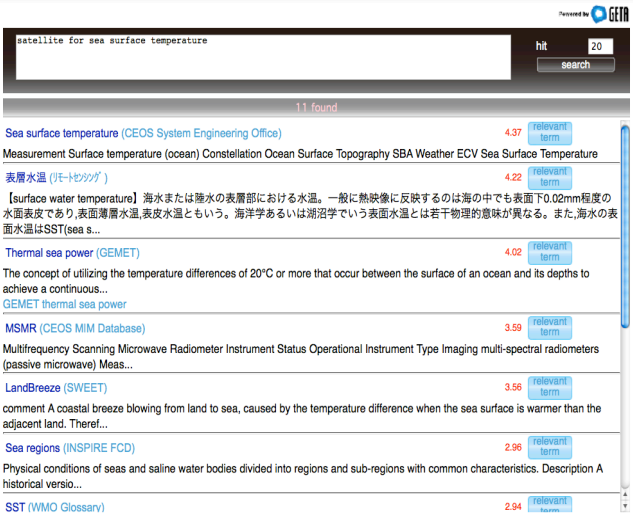

Figure 2. Reverse Dictionary, which retrieve multi data sources

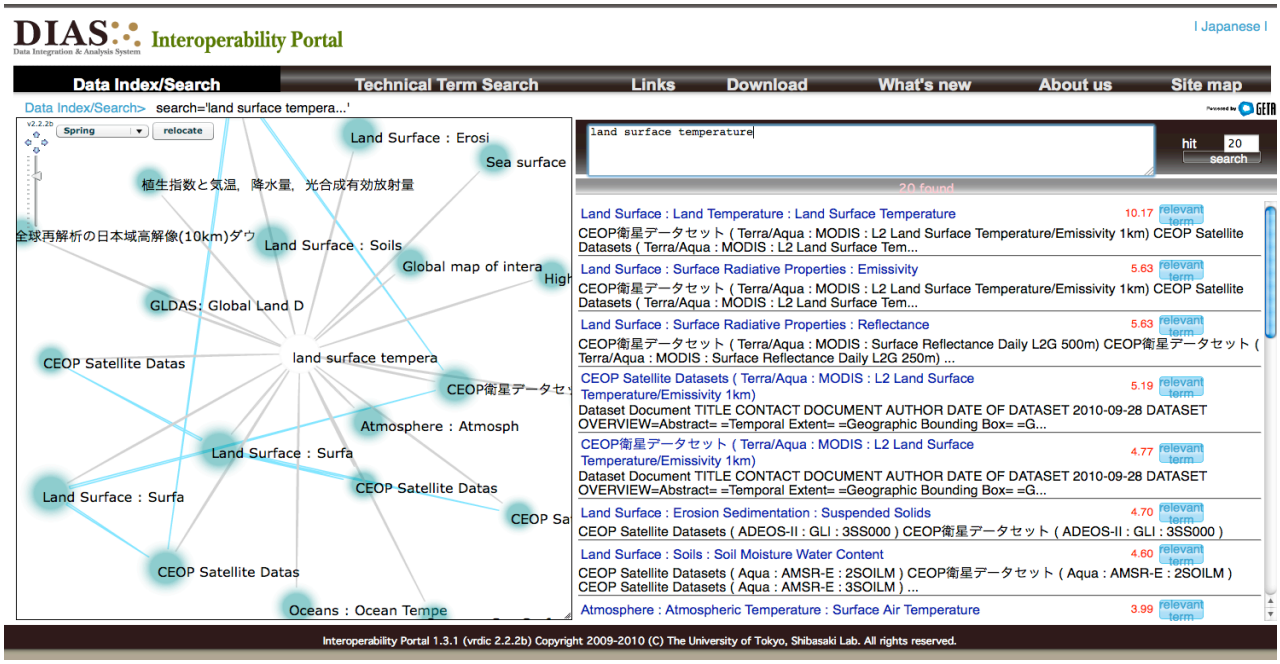

Figure 3. Graph representation

reverse dictionary returns the answer as a list of terms with similarity scores as shown in Figure 2, such as "sea surface temperature" and "MSMR (Multi-frequency Scanning Microwave Radiometer)" in CEOS terminology, "Thermal sea power" in GEMET terminology, and so on. The reverse dictionary relates data by calculation of similarity by using a definition. The user without basic knowledge can discover that a "MSMR" instrument is good for monitoring sea surface temperature and that sea surface temperature is related to "Thermal sea power".

4.3 Graphical Representation

In order to compare associations among the different key words from various ontologies and terminologies which are managed by each Wiki, graph representation as shown in Figure 3 is useful. The graph representation is developed by KeyGraph that is open source of Java library. XML data that is constructed in the Wiki is visualized with the result of information retrieval by the reverse dictionary. All the related terms from various ontologies and terminologies are represented at once.

One of the examples is graph representation is a term from landuse classification schema in Thailand and Indonesia as shown in Figure 4. The term "water body" landuse class can be found in both countries. Apparently, both landuse classes are the same, but the level of hierarchy is a bit different in each classification schema. In the case of Indonesian landuse, "water body" does not include watercourses, but "water body" in Thailand includes all water-related geographical features. Consequently, graph representation proves a clear distinction between the two terms. Then, the new information such as the relations of "water body" in both countries can be created that "water body" class in Thailand is the same as "water" class in Indonesia. This kind of information is treated as newly-created ontological information, and is added through the Semantic MediaWiki.

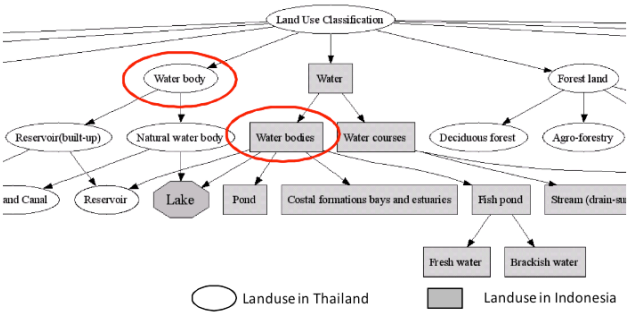

Figure 4. Graph representation for Landuse classification

5. EXPERIMENTAL USE

Experimental implementation has been done under DIAS project. DIAS is a Japanese national key project having missions to archive such earth environmental data and then to analyze global phenomenon through combining and processing these data such as observation data, numerical model outputs, and socio-economic data provided from the fields of climate, water cycle, ecosystem, ocean, biodiversity and agriculture. The aim of DIAS is to share earth observation data and knowledge among different disciplines. Currently, many researchers in the science and engineering fields are participating in DIAS. DIAS is one of GEOSS (Global Earth Observation System of Systems) activities in Japan.

DIAS is tackling a large increase in volume of the earth observation data. DIAS has been developing a core system for data integration and analysis that includes the supporting functions of life cycle data management, data search, information exploration, scientific analysis, and partial data downloading. DIAS is also tackling a large increase in diversity of the earth observation data. For improving data interoperability, DIAS is developing a system for identifying the relationship between data by using ontology on technical terms, and geographical information. DIAS is also acquiring data base information from various sources by developing a cross-sectoral search engine for various databases. Interoperability portal for DIAS has been developed. This portal provides data and metadata search, technical search and

visualization of relations among dataset to very large scale and various earth observation data registered in the DIAS core system.

Most of earth observation data commonly in DIAS have spatial and temporal attributes such as the geographic coverage and the time stamp of data creation with scientific keywords. The metadata standard is published by the geographic information technical committee (TC211) in ISO 19115 and 19139 series metadata standards. From the viewpoint of data users, metadata is useful not only for data retrieval and analysis but also for interoperability and information sharing among experts. In DIAS, document centric metadata registration tool has been developed for reducing time for creating metadata. Since various kinds of datasets stored in DIAS core system increase, it is necessary to support searching datasets based on keywords, spatial conditions, and temporal conditions with created metadata. Datasets are classified into some categories based on such criteria as GCMD Science Keywords or GEOSS SBA (Social Benefit Areas).

Registered metadata and ontological information are managed in the interoperability portal. Dataset is accessed from four categories; persons, places, keywords, and organization. Persons are the responsible person name for dataset. Places are the location of dataset such as country and city name. Keywords are the related keywords that are controlled by ontology registry. And Organization is the information of data provider. The interoperability portal helps keyword retrieval. Figure 3 shows the example of retrieving. The query is "land surface temperature". The result shows not only related dataset but also related researchers name, organization, and available location of dataset with ontological information.

6. CONCLUSION

According to the improvement of observation technologies and earth science studies, a large amount and various kinds of earth observation data including remote sensing data, satellite images and model simulation data are globally being produced by many experts and researchers. At the same time, many kinds of ontology, taxonomies, thesauruses, and gazetteers are being produced in various fields. The ontology registry needs to be developed as a showcase and as a basis for the comparative analysis for better semantic interoperability among diversified earth observation data. The ontology registry is carrying out a component of DIAS and GEOSS interoperability infrastructure.

We have developed the ontology registry and collected the authoritative glossaries, dictionaries, terminologies and ontologies about the earth observation domain and also developed the multi-referential reverse dictionary. The reverse dictionary accepts a group of texts as queries, and return highly related technical terms in the relevance order. Our proposed approach is beneficial for earth observation data interoperability management with ontological information. We will continue to register ontologies and terminologies to this ontology registry. The ontological information can grow autonomously by adding new vocabularies and their relations, becoming more and more useful knowledge.

References from Journals:

Gruber. T.R., 1993, A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220.

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

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

References from websites:

NASA Jet Propulsion Laboratory, California Institute of Technology, Semantic Web fir Earth and Environmental Technology (SWEET).
http://sweet.jpl.nasa.gov/index.html

FAO, AGROVOC Thesaurus.
http://aims.fao.org/website/AGROVOC-Thesaurus/

Global Biodiversity Information Facility (GBIF), Biodiversity Data Portal.
http://www.gbif.org/

U.S. Environment Protection Agency (EPA), Environmental Protection Agency System of Registries.
http://www.epa.gov/sor/

Australian Institute of Health and Welfare, METeOR.
http://meteor.aihw.gov.au/

UK Highways Agency, ITS Metadata Regis-try.
http://www.itsregistry.org.ul/

NASA Goddard Space Flight Center, Global Change Master Directory.
http://gcmd.nasa.gov/Resources/valids/

NetCDF Climate and Forecast Metadata Convention, CF Standard Names.
http://cf-pcmdi.llnl.gov/documents/cf-standard-names/

CEOS Missions, Instruments and Measurements database online.
http://database.eohandbook.com/index.aspx

World Meteorological Organization (WMO), WMO Space Programme – Glossary.
http://www.wmo.int/pages/prog/sat/Glossary.html

European Environment Agency (EEA), GEMET Thesaurus.
http://www.eionet.europa.eu/gemet

Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), CUAHSI's Hydrologic Information System.
http://his.cuahsi.org/

Japan Aerospace Exploration Agency, JAXA Glossary.
http://www.satnavi.jaxa.jp/basic/glossary/index.html

GIS Association of Japan, GIS Glossary.
http://www.gisa-japan.org/

References from Books:

Japan Association of Remote Sensing, 1989, Remote Sensing Dictionary, Kyoritsu Shuppan.