Metadata Management for Integration and Analysis of Earth Observation Data

Akira Takahashi[1] Masashi Tatedoko[1] Toshiyuki Shimizu[1] Hiroko Kinutani[2] Masatoshi Yoshikawa[1]

[1] Graduate School of Informatics,
Kyoto University, Japan
Yoshida-Honmachi, Sakyo-ku, Kyoto 606–8501, Japan
Email: {atakahashi, m.tatedoko}@db.soc.i.kyoto-u.ac.jp,
{tshimizu, yoshikawa}@i.kyoto-u.ac.jp

[2] Earth Observation Data Integration and Fusion Research Initiative (EDITORIA),
The University of Tokyo, Japan
Komaba 4–6–1, Meguro-ku, Tokyo 153–8505, Japan
Email: kinutani@tkl.iis.u-tokyo.ac.jp

Abstract— **Earth observation technologies have developed rapidly during the past few decades. Substantial amounts of earth observation data have been acquired and are currently stored in the literature and databases for various research fields such as climatology, oceanography, agriculture, and ecology. Analyzing and integrating such data might produce valuable data products to promote better understanding of the global environment and to help solve global environmental issues. However, most institutions store and manage their earth observation data independently, with little metadata. Scientists have to struggle to search for valuable data from information outside their research domains and seek uses for these. This paper introduces a conceptual model of earth observation data. Utilizing a model to express earth observation items associated with ontologies, the model is a simple quintuple with information extracted from conventional data models, and it is used to uniquely determine portions of earth observation data, which enables flexible annotations to these data. We also introduce our systems to manage the metadata and user interfaces to encourage users to add annotations to earth observation data that can help scientists discover and understand useful information that can support their research.**

Index Terms— **earth observation data, metadata management, data modeling, ontologies, data lineage**

I. Introduction

Earth observation data have increased both in volume and diversity in recent decades and integrating these data for practical use has attracted a great deal of interest. Earth observation data are collected today by many organizations and institutions from various fields of studies using methods such as in-situ observations, oceanographic observations, remote sensing, weather and climate models, and observations by participating citizens. We should be able to achieve greater understanding and find comprehensive solutions to global environmental issues by integrating such data from different disciplines.

A. Data Integration and Analysis System

The Data Integration and Analysis System (DIAS) project is intended to facilitate multi-disciplined management of earth observation data. It is part of the Global Earth Observation System of Systems (GEOSS), a multinational project for managing earth observation data. Launched in 2006, DIAS is part of the Earth Observation and Ocean Exploration System, which is one of five National Key technologies defined by the Third Basic Program for Science and Technology of Japan. The project was designed to coordinate cutting-edge information science and technology and various research fields examining the earth's environment, to create knowledge enabling us to solve the planet's environment problems and to generate socioeconomic benefits. Several projects within the framework of DIAS have achieved certain applications, such as integrated water resources management, agricultural production management, ocean circulation and fishery resources management. They have also achieved echosystem conservation and participatory monitoring programs.

B. Emerging Problems

More than 100 terabytes of earth observation data were collected from organizations within the DIAS framework and stored in the core system of DIAS during 2007. Several hundred terabytes are to be stored within the next three years. The collection phase of acquiring valuable data has been successful. Nevertheless, scientists must confront several problems that hinder their use of the collected data.

First, most earth observation data have been acquired by organizations and institutions independently; subsequently, the data have been managed in a domain-specific format, intended to be accessed by special application software, and requiring certain labor-intensive efforts on the part of scientists to use the data.

doi:10.4304/jsw.5.2.168-178

As few metadata have been provided, scientists have had to struggle to discover and understand the data that met their demands. However, managing metadata and providing high-quality data products might impose additional burdens on data providers.

To address these problems, we need to establish a model for earth observation data to support interoperability between the data products and to better manage the metadata. We applied the model to metadata provided by the creators of data products. Some geospatial metadata standards [1], [2] already exist to cope with interoperability between spatial data, but no models have been found that use of earth observation data in practice. However, as was described earlier, earth observation data have been rapidly increasing in volume and diversity and unifying metadata is insufficient to enable valuable data to be discovered or understood. Using data annotation and lineage is necessary to support better methods of discovering data and encouraging more research activities.

C. Data Annotation

Management of data annotation should reap further benefits for scientists through their discovery of needed data and deeper understanding of these. Data annotation is ubiquitous on the Internet today. So-called Web 2.0 applications, such as Youtube[1], Wikipedia[2], Facebook[3], and delicious[4] use annotations by users to enhance the value of their content. We believe that this Web 2.0 idea underlying user interactions can also be applied to the region of e-science. User annotations are expected to provide better understanding of data products and might introduce new schemes to enable earth-observation-data products to be evaluated.

Earth-observation data have various content and formats. Remote sensing provides data in images covering wide geographical areas, whereas data from meteorological observations provide temporal sequential data at certain geographical sites. The actual data files might be provided by text formats or binary formats intended to be read by specific applications such as NetCDF [3] or GrADS [4]. We need to produce a conceptual data model that neither relies on data formats nor objects, and that can uniquely determine which data are annotated to discuss how to manage and annotate such data. It must be able to seek a URI for managing earth observation data.

II. Related Works

A. Metadata Modeling

A great deal of work has been done on geospatial metadata modeling. Some metadata have been incorporated into actual data formats, and are hence called *self-describing* formats. NetCDF [3] and GrADS [4], for example, are some major projects on data modeling for software to analyze grid data. Several standards are used for geospatial metadata, such as the Content Standard for Digital Geospatial Metadata (CSDGM) [2], which is used by the Federal Geographic Data Committee. Another standard is that issued by the International Organization for Standards [1].

The ADEPT (The Alexandria Digital Earth Prototype) architecture [5] was proposed by the Alexandria Digital Library (ADL) Project [6] which developed distributed digital library with collections of georeferenced materials. This architecture introduced ADEPT bucket framework to achieve uniform client services which are independent from metadata of heterogeneous items such as image and map in digital libraries. A bucket is an abstract metadata category to which semantically similar source metadata are put together. This aggregation uses manual mappings between metadata fields and buckets. Though we can describe spatial or temporal inclusion relation when we search on the spatial or temporal buckets, there is no discussion about how to integrate heterogeneous data items with different spatial and temporal resolution.

We have introduced a conceptual model for earth observation data, and also introduced a metadata modeling based on the conceptual model. The proposed metadata modeling does not depend on particular data formats. Our main target is numeric data of earth observation and our modeling can uniformly handle earth observation data with different spatial and temporal resolution.

B. Metadata Development Tools

There are several implementations of tools available for developing geospatial metadata, which have been intended to generate standard-compliant metadata. Enraemed[5], MetaD[6], CatMDEdit[7], IME[8], and GeoNetwork[9] represent a brief list of such tools. However, let us take a closer look at IME.

IME The IME (ISO Metadata Editor) is an metadata editor developed in Spain by the Remote Sensing Laboratory of the National Institute for Aerospace Technology (INTA). IME is a tool that features the metadata edition, its modification, and validation according to ISO 19139 and ISO 19115 standards. Fig. 1 has a screenshot of the IME metadata editor. Users can fill out all the metadata defined in the ISO 19115 standard. However, the interface assumes advance knowledge of ISO 19115 standards. It is thus difficult for users of earth sciences to use tools that require understanding of fairly new and unfamiliar standards.

C. Metadata Management Systems

There have been a number of studies on managing data annotations. Some work on managing annotations

[1] http://www.youtube.com/
[2] http://en.wikipedia.org/wiki/
[3] http://www.facebook.com/
[4] http://delicious.com/

[5] http://clearinghouse4.fgdc.gov/enraemed/
[6] http://www.geoportal-idec.net/geoportal/eng/inici.jsp?pag=metad&home=s
[7] http://catmdedit.sourceforge.net/
[8] http://www.crepad.rcanaria.es/metadata/en/index_en.htm
[9] http://sourceforge.net/projects/geonetwork

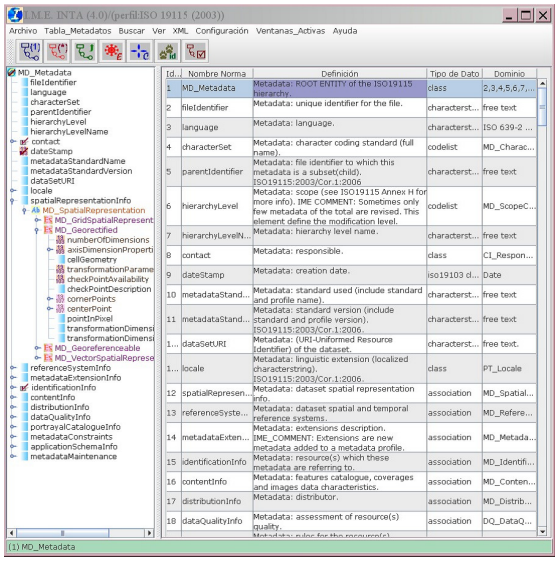

Figure 1. Screenshot of IME.

in HTML documents on the Internet has included Annotea [7] and [8], [9]. Social bookmark services such as delicious and Hatena[10], and other services with user-created content, such as YouTube and Flickr[11] share user annotations and comments to evaluate and classify various content. Annotation-management systems for genomic sequences have also recently been built [10], [11], as well as in the domain of data warehouses and scientific datasets [12]–[15].

We designed and implemented an annotation-management system for earth-observation data. Conventional systems for managing annotations on earth-observation data often attach annotations to all data files, or a bigger granularity, such as the whole dataset. However, users may want to annotate data in several portions of all data files. For example, they may want to annotate data derived from a specific instrument throughout the dataset. We can achieve this by iteratively annotating each data item. However, user semantics can easily be lost by adding another data file to the dataset. We were thus concerned with preserving the user semantics in annotations. To the best of our knowledge, this is the first implementation of a system for managing annotations of earth observation data that preserves user semantics.

The rest of the paper is organized as follows. We introduce our conceptual model for managing annotations of earth observation data in Section III. Section IV discusses several issues concerned with the quality of metadata. Then, we briefly introduce the system we implemented, and present application interfaces to enhance collaboration between scientists in Section V. Finally, Section VI concludes the paper with a discussion of future work.

[10]http://b.hatena.ne.jp/

[11]http://www.flickr.com/

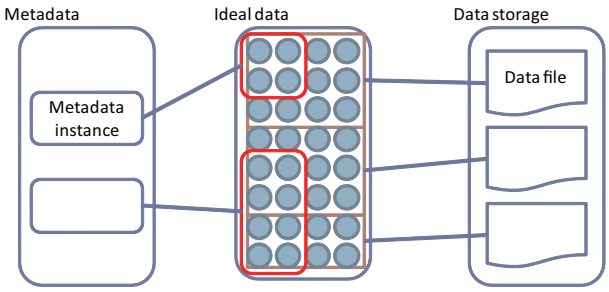

Figure 2. Annotating data with flexible granularity.

III. Annotation Model of Earth Observation Data

The ability to manage annotations with various granularities is necessary to retain user semantics in annotations. Our approach outlined in Fig. 2 was to assume a virtual dataset with fine-grained data, where metadata were annotated against sets of small granules of data. We describe details on how we modelled all granules of earth observation data in the following subsection.

A. Conceptual Model of Earth Observation Data

To refer to a certain portion of earth observation data, we need a conceptual model of data that does not rely on data formats or objects. Consider the example of earth observation data listed in Table I. The data are part of the WMO Resolution 40 dataset provided by NOAA [16]. Two tables exist within the data: a station list for denoting the geographical site of an observatory, and a data file, which indicates the actual value of the data. To refer to the value in the second row and the fourth column (39.8), we need to specify the values of the STATION NAME, YEARMODA, the column name TEMP, as well as the dataset name *WMO Resolution 40*. These four pieces of information—the spatial, the temporal, the observational and the dataset attributes— are general information used to specify earth observation data of any kind.

Using these attributes, we model an earth observation datum, expressed as d, as the following quintuple.

$$d = (ds, s, t, i, v)$$

Each attribute of d describes an aspect of earth observation data. Actually, ds is a dataset identifier, s specifies spatial attributes, and t specifies those that are temporal. In addition, i is the observation item attribute and v denotes the actual value observed (or simulated). We can uniquely determine d, to which the earth observation datum is referring, by using this quintuple. Attributes ds and i play key roles in our model in specifying instances of earth observation data. We further explain each attribute in the following.

1) Dataset Attributes: This attribute denotes the dataset to which the data belong. The value of this attribute is an identifier of a dataset, i.e., it might indicate the source satellite of the data, what climate model was

TABLE I.
EXAMPLE OF EARTH OBSERVATION DATA: WMO RESOLUTION 40

Station list

USAF	WBAN	STATION NAME	CTRY		LAT	LON	ELEV
:	:	:	:	:	:	:	:
477550	99999	HAMADA	JP	JA	34.9	132.067	200
477560	99999	TSUYAMA	JP	JA	35.067	134.017	1470
477590	99999	KYOTO	JP	JA	35.017	135.733	460
477610	99999	HIKONE	JP	JA	35.283	136.25	890
477620	99999	SHIMONOSEKI	JP	JA	33.95	130.933	190
:	:	:	:	:	:	:	:

Data file

STN	WBAN	YEARMODA	TEMP	DEWP	SLP	STP	VISIB	WDSP	MXSPD	GUST
477590	99999	20080101	37.4	25.6	1012.5	1006.7	15.5	3.3	5.1	999.9
477590	99999	20080102	39.8	28.9	1019	1013.2	17.1	3.5	8	999.9
477590	99999	20080103	42.9	28	1020.2	1014.5	21.7	2.7	6	999.9
:	:	:	:	:	:	:	:	:	:	:

used, and which buoy was used. Furthermore, a certain prefix might describe the data processing used.

2) Spatial and Temporal Attributes: The spatial and temporal attributes denote the extent of space and time where the data are valid. The spatial attribute value, s, is a representation of a geospatial point or a region. The temporal attribute value, t, is a representation of the duration or time.

3) Observational Item Attribute: Observation items might describe several characteristics of the data. For example, "max_air_temp" might denote that the observed value is temperature, and is the highest value within a certain period. We define characteristics separately and we define an observation item as a combination of such characteristics. The three characteristics used to define an observation item are:

Target:
The target substance or phenomena of observation, i.e., air, rainfall, or wind.
Property:
The observed property: e.g., temperature, mass, or speed.
Aggregation method:
The method with which the value was aggregated or calculated: e.g., maximum or average values.

The value of the observation item attribute, i, represents a combination of these characteristics. To rigorously determine the characteristics of i, we use ontologies to describe these. Ontologies, such as the SWEET ontologies [17] maintained by the Jet Propulsion Laboratory [18], can provide classes that suits our needs. Fig. 3 outlines the correspondence between characteristics and the ontologies we can use from the SWEET ontologies.

We can describe the observation item using RDF [19] by using ontologies. Fig. 4 has an example of an RDF graph representation of an observation item, `max_air_temp`, the highest value of air temperature measured.

The prefix *ex* used in Fig. 4 denotes the namespace of the ontology used in the DIAS project, which imports the

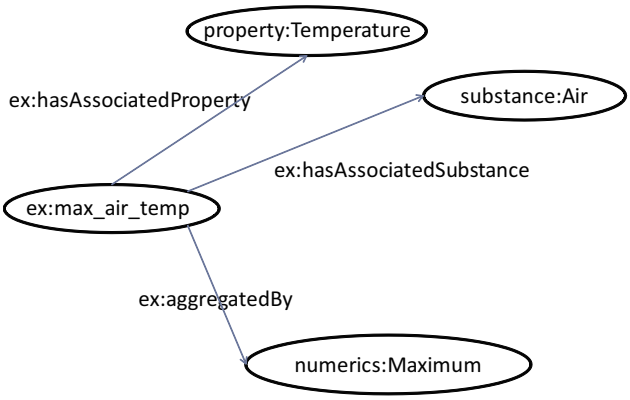

Figure 4. Observation item instance: max_air_temp.

SWEET ontologies and extends their vocabulary.

4) Value Attributes: The value attribute, v, is a representation of the actual value observed, simulated or calculated data, and its unit of measurement, if it exists. v might represent values such as directions or weather, as well as scalars. In addition, null values might be used to indicate missing values.

5) Example: There is an example of earth observation data in Fig. 5, as derived from the values in Table I. We used ISO standards to describe the spatial and temporal attributes. However, this is merely an example. We have no intention of specifying how to implement the descriptions of the attributes. The value of the observation item attribute *mean_air_temp* represents the average air temperature for the day.

Each data item in our data model bears spatial, temporal, and observation item attributes. However, earth observation data are provided as a dataset in most cases. In this subsection, we discuss how to treat a set of earth observation data, viz., an earth observation dataset. We describe an earth observation dataset, D, as

$$\begin{aligned} D &= \{d_1, ..., d_n\} \\ &= (DS, S, T, I, V) \end{aligned}$$

Here, DS, S, T, I, V represents a set of values of all

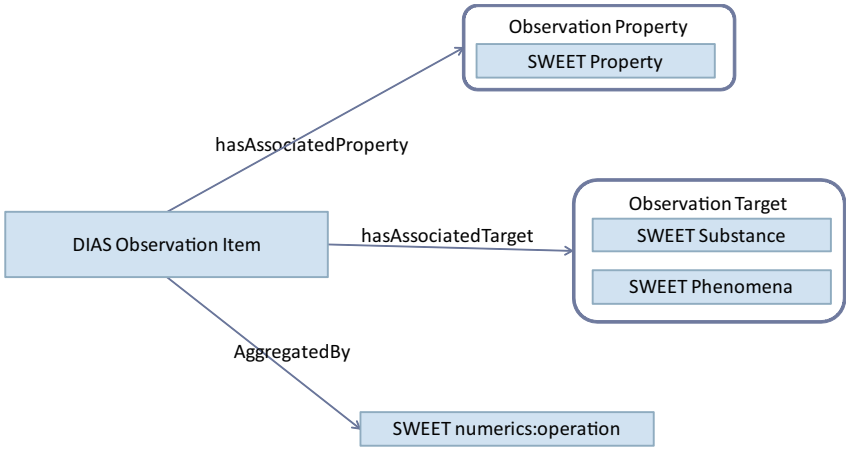

Figure 3. Use of SWEET ontologies.

```
(
  wmoresolution40,
  +35.017+135.733.
  2008-01-01T00:00:00+09:00/
     2008-01-02T00:00:00+09:00,
  mean_air_temp,
  37.4 F
)
```

Figure 5. Examples of data.

attributes, and is defined as

$$DS = \{ds_1, ..., ds_n\}, S = \{s_1, ..., s_n\},$$
$$T = \{t_1, ..., t_n\}, I = \{i_1, ..., i_n\}, V = \{v_1, ..., v_n\}$$

The following defines some properties an earth observation dataset has.

Definition 1: Let D be an earth observation dataset, and i be the value of an observation item attribute. The spatial extent of i in D is the minimum bounded rectangular region that includes all values of the *spatial attribute* of data included in D and that has i as an observation item. This is denoted as $|S_D(i)|$.

Definition 2: Let D be an earth observation dataset, and i be the value of an observation-item attribute. The *temporal extent* of i in D is the shortest duration that includes all values of the temporal attribute of data included in D and has i as the observation item. This is denoted as $|T_D(i)|$.

The next two values are only defined in specific cases.

Definition 3: Let D be an earth-observation dataset, and i be the value of an observation item attribute. If all the values of the spatial attribute of data included in D and has i as the observation item are in the same shape and area, we call the shape and size the *spatial resolution* of i in D and denote this as $\lambda^s_D(i)$.

Definition 4: Let D be an earth observation dataset, and i be the value of an observation item attribute. If all the values of the temporal attribute of data included in D and has i as the observation item have the same length, we call the length the *time cycle* of i in D and denote this as $\lambda^t_D(i)$.

As you can see in the models NetCDF, or HDF-EOS, supports, there are three principal types of geographic distributions of earth observation data, *point, grid,* and *swath*. When data in D which have i as the observation value are distributed in the form of a *grid*, $\lambda^s_D(i)$ is defined as:

$$\lambda^s_D(i) = (|lat_i|, |lon_i|)$$

where $|lat_i|$ and $|lon_i|$ correspond to the length of the zonal and meridional edges of the spatial resolution. In this case, we can define the order for spatial resolution as

$$\lambda^s_D(i_0) \geq \lambda^s_D(i_1) \iff$$
$$|lat_{i_0}| \geq |lat_{i_1}| \wedge |lon_{i_0}| \geq |lon_{i_1}|$$
$$\lambda_s(i_0) = \lambda_s(i_1) \iff$$
$$|lat_{i_0}| = |lat_{i_1}| \wedge |lon_{i_0}| = |lon_{i_1}|$$

If the spatial resolution or the time cycle is common in every i of D, we simply denote them as λ_s, λ_t.

From the discussion at W3C, annotation can loosely be defined as: [20].

> Any object that is associated with another object by some relation.

Annotations on the Web today occur in various forms. They can be RDFs, simple notes or comments, or a number of stars to express user preferences. Although these definitions might be acceptable taking into consideration most content on the Web, we need to slightly alter the definition to discuss annotations for earth observation data. Annotation within our model is defined as *the relation between earth observation data and annotation data*.

We model an annotation datum, denoted as a, as the following triple.

$$a = (u, t, c)$$

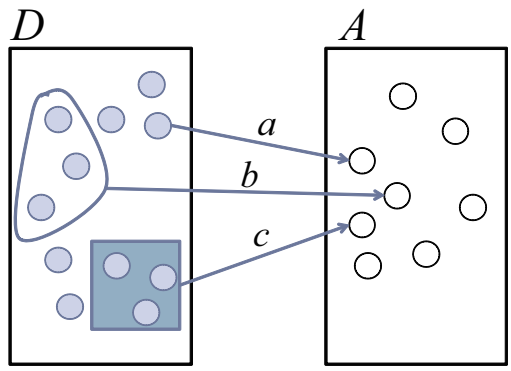

Figure 6. Models of annotations.

Here, u describes a user who made the annotation, t is the valid time the annotation was made, and c represents its content. We denote an annotation, A, of a set of earth observation datum D with annotation datum a as follows.

$$A = (a, D)$$

There is a demand among data users to aggregate earth observation data and annotate them at once when annotating earth observation data. Fig. 6 outlines models of such annotations. In the figure, D and A correspond to conceptual vector spaces where earth observation data and annotation data are denoted as points. Here, annotations a, b, and c represent the three general types of annotations.

Annotation a represents a single data annotation whose subject of annotation is an instance of earth observation data. We denote such annotations as follows.

$$A = (a, (DS, S, T, I, V))$$

Annotations b and c are annotations whose subject is a region in vector space D. We allow two methods to denote these kinds of annotations. The first ignores various dimensions of the earth observation dataset. For example, we might want to annotate all data with the same dataset, temporal, and observation attributes. To meet this requirement, we use an asterisk to represent "do not care". Such annotations with the do not care attribute are written as follows.

$$A = (a, (ds, *, t, i, *))$$

The second method is to use comparison expressions to determine the subset of a dataset. We allow the use of selection conditional expressions, which are defined as follows, in the annotation.

Definition 5: When $X(\in \{DS, S, T, I, V\})$ is an attribute of dataset D, Y is a set of the value of X, and $\theta \in \{\in, \notin\}$ denotes a membership operator. Here, $X\theta Y$ is a conditional clause of D. In addition, when $X(\in \{S, T, V\})$ is an attribute of dataset D, y is a value constant, and θ is a binary operation in the set, $\{<, >, \geq, \leq, =, \neq\}$. $X\theta y$ is also a *conditional clause* of D. The *selection conditional expression* is defined as shown below.

1) a conditional clause of D is a conditional expression of D.

```
A = ( ( Akira Takahashi,
 2007-08-31T15:00:00+09:00,
<iso:CI_ResponsibleParty>
  <iso:individualName>
    Akira Takahashi
  </iso:individualName>
  ...
</iso:CI_ResponsibleParty>
),
```

$(ds, *, *, *, *))$

Figure 7. Annotation with XML syntax.

2) $\neg l$ is a conditional expression of D when l is a conditional expression of D.
3) $l_1 \wedge l_2$ is a conditional expression of D when l_1, l_2 are conditional expressions of D.
4) $l_1 \vee l_2$ is a conditional expression of D when l_1, l_2 are conditional expressions of D.

Let us give an example of an annotation in which information is annotated to data with ds, s, and i corresponding to dataset, spatial, observation item attributes, and where temporal attributes that represent durations after date X.

$$A = (a, (ds, s, t \geq X, i, *))$$

We will give a practical example of an annotation to improve understanding.

A = ((Akira, 2008-08-31T15:00:00+09:00, *systematic error*), (ds,*, $t < 1990 - 01 - 01$, i $\in$ {air_temperature,precipitation },*))

This annotation denotes that there a systematic error in data where their dataset attribute is ds, and the observational items are air temperature and precipitation measured at any location before January 1, 1990.

We specified no syntax that annotation content might have in our data model; generally, no restrictions defined what users could annotate. However, it might be useful if the annotations were available in a machine-readable format. Additionally, we might want to specify the semantics of annotations to distinguish them and avoid mutually exclusive ones. Therefore, we used XML syntax as the annotation content. Using well-known schemas to mark up annotations might increase their interoperability. Let us present an example of an annotation using markups with classes defined using ISO 19115 metadata standards [1] in Fig. 7.

IV. Maintaining Quality of Metadata

This section discusses several issues concerned with the quality of metadata, and describes additional features of the system.

A. User Account Information Management

Previous work on generating metadata in the discipline of earth science has been done by either domain specialists who are technically literate scientists or metadata specialists with profound knowledge of metadata

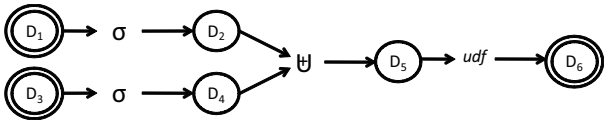

Figure 8. DAG representation of data lineage.

standards. However, Web 2.0 applications such as social bookmarking services involve the general public in generating metadata. Therefore, spam may considerably degrade the quality of metadata if no countermeasures are taken [21]. Despite this, we believe clarifying the accountability of annotations can guarantee the quality of metadata. If we can know what kinds of people created the annotations, we may possibly regard their information as being trustworthy.

As the DIAS system currently requires members in the project to log on, each account is associated with individuals who are professionally affiliated. By disclosing such information, users can determine whether the metadata is trustworthy or not.

However, they may be reluctant to provide annotatations under such circumstances. To avoid this, functions to control access to annotations should be added.

B. Data Lineage

The ability to retrieve the lineage may also contribute to deeper understanding of earth observation data. This section discusses how data lineage can be modeled within the proposed earth observation data model. The lineage of an earth observation dataset consists of a directed acyclic graph (DAG) that consists of dataset nodes and operation nodes. There is an example of lineage data outlined in Fig. 8.

The circle nodes in Fig. 8 are the dataset nodes representing earth observational datasets. We have distinguished datasets that are actually stored in a system (indicated by nodes with concentric circles) from temporal datasets (indicated by single-circled nodes). The process designated in the figure involves three steps:

1) Select subsets from D_1, D_3 and obtain temporal datasets D_2, D_4,
2) Merge D_2, D_4 and obtain D_5, and
3) Aggregate D_5 in a user-defined method and generate D_6.

This section formalized the transformation of earth observation datasets and introduced the data lineage model.

1) Transformation of Datasets: We have discussed how to describe datasets in our conceptual model. To describe lineage data, we must now define how to describe the process of generating datasets. We use a modification of relational algebra to describe transformations of datasets, and use them to describe the processing history of datasets.

Set operators Two of the six basic operators are useful for manipulating the earth observation dataset: the set union and the set difference. Unlike relational algebra, a Cartesian product is not necessary in our model because the number of attributes is constant.

Selection The definition for selection operation has been given in Section III.

Conversion We define conversion as an operation that only involves changes in the observational value attributes. Conversion describes transformation of datasets such as the calibration of raw values or a change in units. The conversion operation for dataset D is defined as explained below.

Definition 6: When u_1, u_2 are units of measurement and $f(v)$ gives a value measured as u_2, which is semantically equivalent to value v measured as u_1, the *conversion operation* for dataset D is expressed as $\varepsilon_{u_1 \to u_2}(D)$ or $\varepsilon_{f(v)}(D)$, and is defined as

$$\begin{aligned}\varepsilon_{u_1 \to u_2}(D) &= \varepsilon_{f(v)}(D) \\ &= \{(ds_i, s_i, t_i, i_i, v_i') | (ds_i, s_i, t_i, i_i, v_i) \in D \\ &\quad \wedge \text{ unit of } v_i = u_1 \wedge \text{ unit of } v_i' = u_2 \\ &\quad \wedge \text{value of } v_i' = f(\text{value of } v_i)\}\end{aligned}$$

Aggregation An operation with a change involving either the observation period, the spatial region, the spatial resolution, or the time cycle of a dataset is defined as an *aggregation operation* in this research. It is respectively as for the observation period of the set of data with observation item i of dataset D and the range of the space, the space resolution, and the time cycle. The aggregation operation is expressed as

$$opt_{i,|S(i)|,|T(i)|,\lambda^s(i),\lambda^t(i)} D$$

When D has spatial resolution $\lambda_D^{s_0}(i)$ or time cycle $\lambda_D^{t_0}(i)$, the following constraints can be applied to the value of $\lambda^s(i), \lambda^t(i)$:

$$\lambda^s(i) \geq \lambda_D^{s_0}(i), \lambda^t(i) \geq \lambda_D^{t_0}(i)$$

Therein, opt describes the method of aggregation; it is a member of the set, max, min, sum, $average$, and $count$, which correspond to operations that return the maximum values for observation, minimum value, summation, the arithmetic mean, and the number of instances. Any other operation that involves changes in the spatial attributes or the temporal attributes is expressed as $aggr$.

Aggregation operation in some cases might involve temporal periods or spatial regions without a data instance. For instance, consider an operation for taking the mean temperature in September in a dataset where the observations started from September 15. In such cases, we assume that there are data instances with ımissing values for the value of observation attributes before September 15.

2) Merge: We define ımerge as a binary operation, whereby two datasets D and D' are the its input. We need to specify four parameters to use the operator.

1) Aggregation method opt
2) Spatial resolution λ^s
3) Time cycle λ^t
4) Observation item i

The merge operation is

$$D \uplus_{opt,\lambda^s,\lambda^t,i} D'.$$

Similar to the join operation in relational algebra, *merge operation* is a syntax sugar; it is a representation of the following operations.

1) Select data from D and D', where the observation item attribute has the same *target* and *property* with i
2) Aggregate all datasets by using the operation denoted by opt and set their time cycle to λ^s and their spatial resolution to λ^t, i.
3) Return the union of the aggregated datasets.

Natural Merge We might lack some parameters in merge operation. In such cases, the parameters will be set automatically to a value derived from the two inputs, D and D'. The method of choosing the value is explained in what follows.

- **When no opt, i parameter is set:** When data with i as the observation item exist in D and data with i' as the observation item exist in D' and i and i' shares the same target and property, and i and i' both have the same aggregation method, or when one of them has no aggregation method, then opt is the method used to aggregate these items. Each method is used to aggregate each dataset if they have different methods of aggregation. In any other case, opt will not be determined, and merge operation will return an empty set.
- **When no λ^s, λ^t, i parameter is set:** When data with i as the observation item exist in D and data with i' as the observation item exist in D', and i and i' share the same target and property, λ_s and λ_t will be the lowest common multiple of the spatial resolution or time cycle of i in D and i' in D'.

We designate a merge operation with no parameter as a natural merge and describe it as shown below.

$$D \uplus D'$$

Mergeability Many data integrations involving two datasets can be described by using merge operation. Using merge operation, we can define mergeability:

Definition 7: When D and D' are earth observation datasets, and

$$D \uplus D' \neq \phi$$

where ϕ denotes an empty set, we say that D is *merge-able with* D'.

With this definition and the definition of natural merge operation, we obtain the following theorem.

Theorem 1: When D and D' are earth observation datasets, and there are data with i as the observation item in D and there are data with i' as the observation item in D', and i and i' share the same target and property, and at least one of the two items has a method of aggregation, we say that D has a possibility of merging with D', or that D is merge-able with D'.

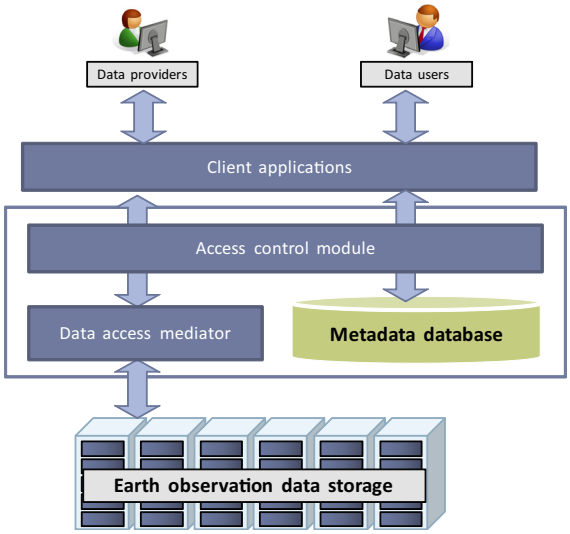

Figure 9. System overview.

Using this concept of mergeability, we can offer datasets to users that can be used for integrated analysis of earth observational datasets.

V. Plan for Implementation of System

This section explains the system we implemented, and introduces some application interfaces for annotating metadata.

A. Overview of Our System

Fig. 9 presents an overview of our system. The **data access mediator** provides the mapping between our conceptual earth observation datum model and the actual data model used in the underlying storage. Users can retrieve data by specifying the conditions for the quintuple, with no awareness of the data model schema used in the actual data storage. As there are various data formats used in the DIAS project, schema mapping is currently created manually. Automatic processing of schema mapping is work we wish to explore in the future.

Many earth observation data products have restrictions on their use, and users may want to control access to metadata they have created. Therefore, we have incorporated an access control module to manage user accounts and how they can access each data. Client software and applications can be built on this access control module.

B. Application Interfaces

This section describes some implementations of user interfaces. Users can obtain supplementary information on datasets by utilizing metadata, and encourage further understanding of these datasets. However, we must motivate users to enter metadata to obtain enough information. This is a difficult challenge in designing user interactions. If users are not aware of how they can benefit from entering metadata, no one will provide metadata, and the system

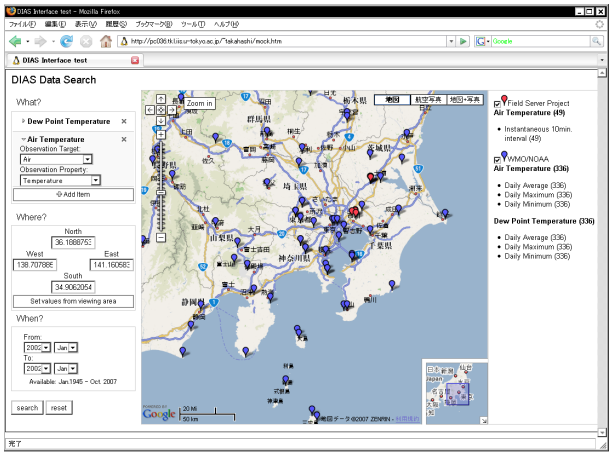

Figure 10. The data retrieval interface.

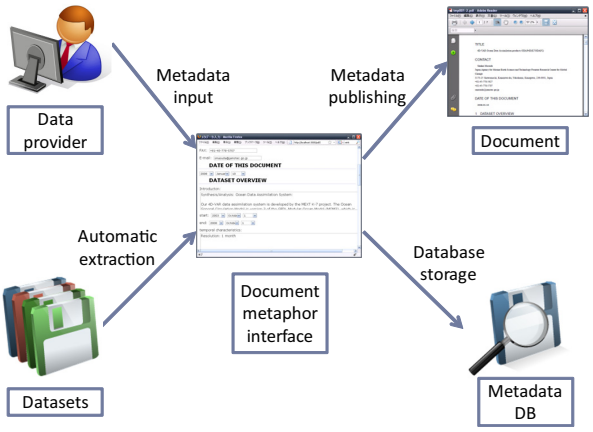

Figure 11. Document metaphor interface.

will be forced to a halt. Our policy in designing user interfaces is to show how effective metadata annotation is as quickly as possible.

1) Data Retrieval Interface: We previously implemented a data retrieval system [22] for earth observation data. The policy in designing the interface was to effectively combine searches from three aspects: items, regions, and times. In addition it aims to achieve a sight grasp of the data with availability of data shown on a map-based interface.

Figure 10 has a screenshot of the Web-based interface we developed. The left column is used to input queries, and the right column summarzes information on the query. The center column is a map interface using GoogleMapsAPI [23], and sites corresponding to the query will be visualized on the map.

You can specify the observation item, region, and period in the query input part to narrow down users' results. The item is indicated by specifying the observation target, properties, and the interval. Users can specify two or more observation items, and also determine whether the items should be contained in the result site (AND retrieval) or at least one of the alternatives is contained (OR retrieval). We can search the observation point within a rectangular area obtained by specifying the upper bound's and the lower bound's latitude and longitude. When either of three parameters is specified, the alternatives for the other parameters are automatically limited, according to the specified value, to ensure we can at least obtain one site as a result. The system will automatically query the database to prefetch results and provide the summary on the right and the center column. As users can easily see what kind of data are available and to what extent, their efficient behavior in retrieval is supported.

We are planning to incorporate an annotating function into this interface. We can see the query result in the center map column of the interface with the prefetching function. By simply clicking a marker representing a result, we may see all annotation data, annotated to the particular dataset in the column at right. Users can then evaluate the annotation, choose to download the data, or annotate another information item to the corresponding dataset.

2) Interface for Annotating Metadata with Document Metaphor: Since data providers often understand the data the most, large portions of structured metadata are generated by them. However, conventional metadata publishing tools, such as GeoNetwork[12] and NOAA ArcView Extension[13], require a profound understanding of metadata schemas. We have provided a document metaphor annotation interface to avoid having to study metadata schemas in detail. Data providers often manage and publish documents to explain the datasets they have produced. By providing the tools to generate such documents within our framework, users can save the trouble of having to describe both document and input metadata. Fig. 11 overviews the interface's functions. The document metaphor annotation interface is a form-style editor for metadata. Users can access the interface through conventional Web browsers. The interface provides the section titles of the document, and users fill out the content of all sections (Fig. 12). If the dataset the user is going to refer to is described in NetCDF or GrADS formats, some portions of the sections can be filled out with metadata extracted from the data. Also, some of the information on the data provider may be automatically filled out, if the user information is registered in advance. Inputs are stored in databases, and users can either download the document in Portable Document Format (PDF), or publish the document in HTML.

VI. Summary and Future Work

We proposed a conceptual data model for annotating earth observation data. Utilizing the conceptual model enables users to state metadata without having to be concerned with data models used in actual data storage, and preserve the user semantics of the annotations. We also introduced our system to manage the metadata. We

[12]http://geonetwork-opensource.org/
[13]http://www.csc.noaa.gov/metadata/download.html

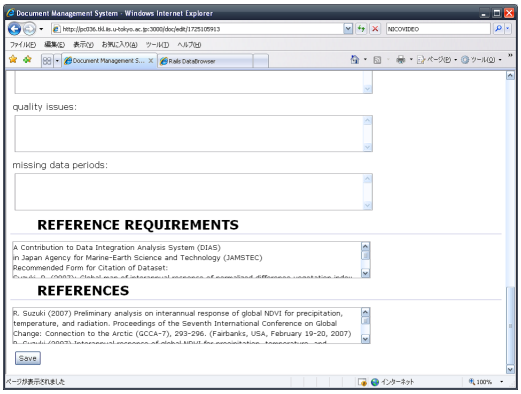

Figure 12. Metadata registration interface.

are currently implementing additional user interfaces for utilizing metadata. Our future work includes collaboration by users in annotation, and management of data provenance information within our framework.

ACKNOWLEDGEMENTS

This research was supported by the "Data Integration and Analysis System" funded by the National Key Technology, Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

[1] International Organization for Standardization, "ISO 19115:2003, geographic information metadata."
[2] Federal Geographic Data Committee, "Content standard for digital geospatial metadata. FGDC-STD-001-1998," June 1998.
[3] "NetCDF(Network Common Data Form)," http://www.unidata.ucar.edu/software/netcdf/.
[4] "Grid Analysis and Display System (GrADS)," http://www.iges.org/grads/.
[5] G. Janee and J. Frew, "The ADEPT digital library architecture," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Portland, USA, June 2002, pp. 342–350.
[6] L. L. Hill, G. Janee, R. Dolin, J. Frew, and M. Larsgaard, "Collection metadata solutions for digital library applications," *Journal of the American Society for Information Science*, vol. 50, no. 13, pp. 1169–1181, 1999.
[7] J. Kahan, M.-R. Koivunen, E. Prud'hommeaux, and R. R. Swick, "Annotea: an open RDF infrastructure for shared web annotations," *Computer Networks*, vol. 39, no. 5, pp. 589–608, 2002.
[8] D. LaLiberte and A. Braverman, "A protocol for scalable group and public annotations," *Computer Networks and ISDN Systems*, vol. 27, no. 6, pp. 911–918, 1995.
[9] M. A. Schickler, M. S. Mazer, and C. Brooks, "Pan-browser support for annotations and other meta-information on the world wide web," *Computer Networks*, vol. 28, no. 7-11, pp. 1063–1074, 1996.
[10] "biodas.org," http://biodas.org/.
[11] R. D. Dowell, R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein, "The distributed annotation system," *BMC Bioinformatics*, vol. 2, p. 7, 2001.
[12] Y. Cui and J. Widom, "Lineage tracing for general data warehouse transformations," *VLDB J.*, vol. 12, no. 1, pp. 41–58, 2003.
[13] P. Buneman, A. Chapman, and J. Cheney, "Provenance management in curated databases," in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, Chicago, USA, June 2006, pp. 539–550.
[14] R. Bose, "A conceptual framework for composing and managing scientific data lineage," in *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, Edinburgh, UK, July 2002, pp. 15–19.
[15] "Trio, A System for Integrated Management of Data, Uncertainty, and Lineage," http://infolab.stanford.edu/trio/.
[16] "National Oceanic & Atomspheric Administration," http://www.noaa.gov/.
[17] "Semantic Web for Earth and Environmental Terminology (SWEET)," http://sweet.jpl.nasa.gov/ontology/.
[18] "Jet Propulsion Laboratory (JPL)," http://jpl.nasa.gov/.
[19] "Resource Description Framework (RDF)," http://www.w3.org/RDF/.
[20] "Collaboration, Knowledge Representation and Automatability," http://www.w3.org/Collaboration/.
[21] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, "Can social bookmarking enhance search in the web?" in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Vancouver, Canada, June 2007, pp. 107–116.
[22] A. Takahashi, H. Kinutani, and M. Yoshikawa, "Research for an integrated earth observation data retrieval framework based on metadata development with ontologies," in *Proceedings of the IPSJ Symposium on DataBase and Web (DBWeb2007)*, 2007, (in Japanese).
[23] "Google Maps API," http://www.google.com/apis/maps/.

Akira Takahashi received a B.E. degree in Engineering and an M.E. degree in Informatics from Kyoto University in 2007 and 2009, respectively.

He is currently a system engineer at NS Solutions Corp.

Masashi Tatedoko received a B.E. degree in Information Science from Osaka Institute of Technology in 2008.

He is currently pursuing an M.E. degree in Informatics at Kyoto University.

Toshiyuki Shimizu received a B.E. degree in Engineering in 2003, an M.E. degree in Information Science from Nagoya University in 2005, and a Ph.D. degree in Informatics from Kyoto University in 2008.

He is currently an assistant professor at Graduate School of Informatics, Kyoto University. His current research interests include XML databases, indexing techniques, and information retrieval. He is a member of ACM.

Hiroko Kinutani received a B.S. degree in Mathematics in 1976, an M.E. and Ph.D. degree in Information Science from Nara Institute of Science and Technology in 1997 and 2002, respectively.

She is currently a specially appointed research associate at Earth Observation Data Integration and Fusion Research Initiative, The University of Tokyo. Her current research interests include XML databases and metadata. She is a member of ACM and IPSJ.

Masatoshi Yoshikawa received a B.E., an M.E. and a Ph.D. degree in Information Science from Kyoto University in 1980, 1982 and 1985, respectively. He was on the faculty of Kyoto Sangyo University from 1985 until 1993. From 1989 to 1990, he was a visiting scientist at Computer Science Department, University of Southern California. In 1993, he joined Nara Institute of Science and Technology as an Associate Professor of Graduate School of Information Science. From April 1996 to January 1997, he has stayed at Department of Computer Science, University of Waterloo as a visiting associate professor. From June 2002 to March 2006, he served as a professor at Nagoya University. From April 2006, he has been a professor at Kyoto University.

His general research interests are in the area of databases. His current interests include XML databases and index structures for text and multimedia data. He is a member of ACM, the IEEE Computer Society and IPSJ.