

## 地球観測データ統合解析のためのデータ系譜と アノテーションのモデル化

高橋 慧<sup>†</sup> 絹谷 弘子<sup>†</sup> 吉川 正俊<sup>†</sup>

<sup>†</sup> 京都大学大学院 情報学研究科 〒606-8501 京都市左京区吉田本町

<sup>†</sup> 東京大学 地球観測データ統合連携研究機構 〒153-8505 東京都目黒区駒場 4-6-1

E-mail: †atakahashi@db.soc.i.kyoto-u.ac.jp, ††kinutani@tkl.iis.u-tokyo.ac.jp, †††yoshikawa@i.kyoto-u.ac.jp

あらまし 近年の地球観測技術の進歩に伴い、気候、海洋、農業、生態系、水循環などの関連機関において地球観測データの観測、蓄積が行われている。このような多様な分野にまたがるデータを統合、解析する基盤を構築することによって地球環境への理解を深め、地球環境問題の解決や災害対策に有益な情報の提供が可能となる。本研究ではこのような基盤においてデータ統合、解析を実現するために地球観測データセット及びデータセットに対する情報のアノテーションのモデル化を行い、統合利用のためのデータ操作体系、問合せ体系を提示しそれを利用したデータ系譜情報のモデルを構築する。

### Data Lineage and Anotation for Integration and Analysis of Earth Observation Data

Akira TAKAHASHI<sup>†</sup>, Hiroko KINUTANI<sup>†</sup>, and Masatoshi YOSHIKAWA<sup>†</sup>

<sup>†</sup> Graduate School of Informatics, Kyoto University Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

<sup>†</sup> Earth Observation Data Integration and Fusion Research Initiative (EDITORIA) Komaba 4-6-1,  
Meguro-ku, Tokyo, 153-8505 Japan

E-mail: †atakahashi@db.soc.i.kyoto-u.ac.jp, ††kinutani@tkl.iis.u-tokyo.ac.jp, †††yoshikawa@i.kyoto-u.ac.jp

**Abstract** Technologies for the Earth observation has developed in the past years, and large amount of Earth observation data has been acquired by institutes among various fields of researches, such as climatology, oceanics, agriculture, hydrology, ecology, and so on. Analysis and Integration of such data may provide valuable data products to promote further understanding of the Earth enviroment and solve earth enviromental problems. Systems for such purpose has gained interests. In this paper we propose a conceptual data model for systems for integration and analysis of Earth observational data, metadata annotation model for such data and model for preserving lineage of data products.

#### 1. はじめに

近年、リモートセンシング技術の向上や気象観測における測定機器の観測精度の上昇、シミュレーションのための大規模計算機インフラの整備などの進歩に伴い、気候、海洋、農業、生態系、水循環などの様々な研究分野において観測、蓄積される地球観測データはその質、量共に大幅に増加してきている。このような多様な分野にまたがるデータを統合、解析する基盤を構築することによって地球環境への理解を深め、地球環境問題の解決や地球温暖化防止や豪雨、干ばつなどの被害軽減に有益な情報を提供することが可能となると考えられている。

しかし、気象、水循環、農業や生態系など複数の専門分野に

わたる膨大な蓄積データの中から、統合解析に利用できる有用で高品質なデータを研究者が探しだすのは容易ではない。そこで、データの検索や内容理解、データの品質管理の為にデータに対する柔軟なアノテーションの実現、及びデータプロダクトの再現性、責任の明確化の為にデータプロダクトの生成過程を表現するデータ系譜情報の管理が必要である。また Web 上のデジタルコンテンツに対するアノテーションと同様に、地球観測データの粒度に応じて柔軟なアノテーションを行う枠組みを作ることにより、データの検索等に役立てるだけでなく、データの利用者の評判や、データの利用頻度などの情報を付与することで、従来の地球観測データプロダクトには存在しなかったデータの評価スキームを構築することが期待できる。

そこで本稿では、地球観測データに対するアノテーションとデータ系譜情報保存を実現するためのデータモデルについて紹介を行う。

本稿の構成は以下の通りである。まず2.で本研究で用いる地球観測データのモデリングについて説明し、次に3.でモデルに対するデータのアノテーション方法を提示、4.でデータ系譜情報のモデリングについて説明、5.で関連研究について紹介し、最後に6.でまとめを行う。

## 2. 地球観測データモデル

### 2.1 地球観測データ

我々が扱う地球観測データにはリモートセンシング技術によって取得された衛星画像、地上、海上における気象観測などの観測行為によって得られるデータや、気候モデルにより得られたシミュレーションの結果など多種多様なデータが含まれる。また同じ観測結果であっても単にセンサの電圧を記録したデータや、観測項目に応じた単位変換を行ったデータ、観測結果をそのまま記録したデータと観測毎にその有意性が検証された、品質管理を行ったデータなどが存在する。またこれらのデータのフォーマットは一般的な画像ファイルや、NetCDF [1], GrADS [2] 等の専門アプリケーションでの利用を前提としたバイナリフォーマット、CSV などの ASCII フォーマットなど多岐にわたる。

このような多様なデータを扱う場合、個々の観測値だけでなく、これらのデータは目的に応じてデータプロダクト単位やプロジェクト単位で扱う場合もある。CCLRC (Council for the Central Laboratory of the Research Councils, 現 Science and Technology Facilities Council [3]) では地球観測データのメタデータ管理階層を「分野 (Study), データセット (Dataset), データオブジェクト (Data Object)」の3階層で定義しており [4], これらの3階層に対してそれぞれメタデータをつけて管理することが有用であるとされている。しかし例えば個別の観測値について品質管理を行った場合には Data Object よりさらに細かい単位でメタデータを付与する必要がある。また特定の地域で観測された Dataset 内の複数 Data Object に対してメタデータを付与したい場合なども考えられる (地方自治体の情報など)。

これらの要求に応え、地球観測データにアノテーションする場合、データフォーマットやデータオブジェクトに依らない概念的なモデルが必要である。

地球観測データには一般に観測された空間情報、時間情報、及び観測されたパラメータの意味を表現する観測項目情報などの情報が紐づけられている。そこで我々は地球観測データ (d) を以下の5つ組によって表現する。

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

ds, s, t, i, v はそれぞれデータセット識別子、空間属性、時間属性、観測項目属性、観測された値及びシミュレーションの結果を表わす。この5つ組において、ある地球観測データに関して ds, s, t, i が定まると v が一意に定まるとする。5つ組を利用することにより観測されたデータを一意に定めることができる。各

属性の詳細な説明を以下に述べる。

#### (1) データセット識別子 (ds):

データのまとまりを表す識別子。海洋のブイ ID や衛星の名前、気候モデル名やそのパラメータなどに対応した値を持つ。またそれぞれの4.節で述べる演算を行ったことを示す識別子などを付与し、v が一意に定まる制約を満たす用に識別子を定める。

#### (2) 空間属性 (s):

地球観測データの持つ、地点や地域など地球上での位置を特定する属性

#### (3) 時間属性 (t):

地球観測データの持つ時刻または期間などの時間に関する情報を特定する属性。

#### (4) 観測項目属性 (i):

観測された値がどのような観測項目に該当するか。地球観測データにおいて観測項目を決定付ける要素は様々であり、また観測を行う機関や分野の違いによって、同じ観測項目でも求められる観測条件が異なる場合がある。例としては気温という観測項目があげられる。気温は一般には大気温度の事を指すが、気象分野においては気温は地上から一定の高さで外気に触れない状態で観測されたものを示す。しかし、本研究では利用者の専門外の分野における観測データの発見を目的とするため、このような分野間で異なる観測項目を個別に定義するのではなく、観測項目の本質的な表現にとどめて扱う。以下に本研究で必要と考える観測項目の4つの要素を示す

観測対象: 大気や土壌、降雨など、観測の対象となる物質や現象。

観測周期: 観測が行われる周期や、観測を集計する期間。データが観測時の瞬間的な値を示す場合はこの要素は不要である。

観測物理量: 温度や湿度、質量、速度などの観測対象において観測されている物理量。

集約方法: 平均や最大・最小など、観測項目がある一定の期間のデータとして与えられる場合にどの様に集約・算出されているか。データが観測時の瞬間的な値を示す場合はこの要素は不要である。

観測項目属性を実装する際には [5] で示したように RDF で表現することが考えられる。

#### (5) 観測値 (v):

データの観測された値及びその単位。0 °C や 40% などのスカラー値に加え、方角などの値も持つ。また観測を行わなかった、または行ったが値が取得できなかったことを示す欠損値なども表現する。

以下に地球観測データのインスタンスを示す。

$$d_{ex} = (ex, +350121+1354642, 2008-04-01T12:00:00+09:00, air.temperature, 10.2 \text{ } ^\circ\text{C})$$

このインスタンスはデータセット *ex* における、ある地点のある時刻の瞬間の気温を表わしている。地点及び時刻の表記は実装方法に依存する。

## 2.2 地球観測データセット

前述のデータモデルでは一つの観測値に対して個別に空間属

性、時間属性などを定めるデータモデルであった。しかし、実際の地球観測においては同じ観測項目を特定の地点で連続して観測を行うなど、データセット内でそれぞれの属性の値を共有している場合が多い。CCLRC の 3 階層の一つを構成していることから、一般に地球観測データを扱う場合データセット、つまり同様なデータのまとまりを扱うことが有用であると考えられる。本研究のモデルでは地球観測データセット  $D$  を以下のように表す。

$$D = \{d_1, \dots, d_n\}$$

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

ここで、 $DS, S, T, I, V$  は  $d_i = (ds_i, s_i, t_i, i_i, v_i)$  に関して以下の条件を満たすような各属性の属性値集合である。

$$DS = \{ds_1, \dots, ds_n\}, S = \{s_1, \dots, s_n\},$$

$$T = \{t_1, \dots, t_n\}, I = \{i_1, \dots, i_n\}, V = \{v_1, \dots, v_n\}$$

$D$  内のある観測項目  $i$  の観測の空間範囲、期間をそれぞれ  $|S(i)|, |T(i)|$  と表す。以下、地球観測データセットが一定の制約を満たす場合に定義される値について説明する。地球観測データセット内のある観測項目におけるデータは一定の時間周期で観測されている場合がある。地球観測データセット  $D$  において、観測項目  $i$  をもつデータの  $T$  が一定の時間間隔を置いて分布している場合、この時間間隔を時間周期と呼び、 $\lambda_t(i)$  と表す。

一般に、地球観測データの空間分布は任意の観測点からなる point 型の分布、 $z$  観測衛星の軌道方向に垂直/水平な格子軸を持つ swath 型の分布、緯度、経度方向にそれぞれ水平な格子軸を持つ grid 型の分布の三種類が存在する。地球観測データセット  $D$  に属するある観測項目  $i$  を持つデータが全て grid 型の分布を持つ時、観測の空間解像度を  $\lambda_s(i)$  と表す。 $|lat_i|, |lon_i|$  をそれぞれ観測項目  $i$  を持つデータの緯度、経度における間隔とする時、 $\lambda_s(i)$  を以下のように定義する

$$\lambda_s(i) = (|lat_i|, |lon_i|)$$

$$\lambda_s(i_0) \geq \lambda_s(i_1) \iff$$

$$|lat_{i_0}| \geq |lat_{i_1}| \wedge |lon_{i_0}| \geq |lon_{i_1}|$$

$$\lambda_s(i_0) = \lambda_s(i_1) \iff$$

$$|lat_{i_0}| = |lat_{i_1}| \wedge |lon_{i_0}| = |lon_{i_1}|$$

$D$  に属するデータの空間解像度、時間間隔が観測項目に依らず一定である場合、それぞれを単に  $\lambda_s, \lambda_t$  と表わす。

### 3. データに対するアノテーション

地球観測データを利用する際には有用なデータの発見やデータの性質の理解のために地球観測データに対して何らかの情報アノテーションし、役立てることができる。地球観測データ

に対するアノテーションではそのアノテーションがどの範囲のデータに対して有効なのかを明確に記述する必要がある。本章では提案する地球観測データモデルに対してどのようにアノテーションの影響範囲を表記するかを論じる。

地球観測データに対するアノテーションにおいて、最も単純なアノテーションは、地球観測データの 5 つ組の値を全て指定する場合である。この場合のアノテーションを以下のように記述する。

$$(ds, s, t, i, v) \longrightarrow annotation$$

またデータセット  $D$  全体に対するアノテーションは以下のように記述する

$$D \longrightarrow annotation$$

実際にアノテーションを行う場合は一定の領域に所属するデータの集合に対して一括して情報をアノテーションしたいという要求が考えられる。本モデルではこの要求にこたえるためにドントケアの使用、選択条件指定を可能にする。以下に、空間属性および観測項目属性にドントケアを利用したアノテーションの例を以下に示す。

$$(ds, *, t, i, *) \longrightarrow annotation$$

この場合アノテーションはデータの空間属性および観測値に関わらず、データ識別子に  $ds$ 、時間属性に  $t$ 、観測項目情報に  $i$  を持つ全てのデータに対して有効である

また時間属性に選択条件指定を用いたアノテーションは以下のように表記される

$$(ds, s, t \theta X, i, *) \longrightarrow annotation$$

具体的なアノテーションの例を示す。

$$(ds, *, t < 1990-01-01, \{air\_temperature, precipitation\}, *) \longrightarrow systematic\ error$$

この例ではデータセット  $ds$  に所属する 1990 年より前に観測された任意の地点の気温及び降水量を観測したデータに関して系統誤差が存在すること表わしている。また例でのアノテーションは概念的なものであり、実装の際には XML タグを用いた構造化などの工夫が必要である。

#### 3.1 アノテーションの継承

地球観測データ統合解析基盤においてデータのアノテーションを行う場合、あるデータセット全体に対してアノテーションを行い、さらに局所的に別の、最初のアノテーションとは背反した情報をアノテーションを行いたいという要求が考えられる。この場合、最初につけたアノテーションに手を加え、新しい対象とそれ以外のデータにアノテーションを細分化するアプローチが考えられるが、この手法を採用するとアノテーションを行う度にアノテーションが細分化され個々に管理する必要がある。また最初のアノテーションがデータセット全体を対象にしているというセマンティックが保存されなくなってしまう。

そこで、この問題を解消するために、本モデルではアノテ

ション対象となるデータセットの一方が他方の真部分集合である場合、前者のアノテーションを優先して採用する事にする。以下に例を示す

- A:  $(ds, *, t < 1990-01-01, \{air\_temperature, precipitation\}, *)$   
 → systematic error  
 B:  $(ds, *, t < 1970-01-01, air\_temperature, *)$   
 → no systematic error

この場合 A のアノテーションによってデータセット  $ds$  に所属する 1990 年より前に観測された任意の地点の気温及び降水量を観測したデータに関して系統誤差が存在することが表現されているが、その後 1970 年より前の気温データの系統誤差が修正された場合、B のアノテーションを新たに付与することにより、A のアノテーションを修正せずとも 1970 年より前の気温データに系統誤差が存在しないことを表現できる。

なお本研究ではどちらかのデータセットが一方の真部分集合になっていない場合はどちらのアノテーションを用いるかは定めない。

### 3.2 アノテーションに対する問合せ

アノテーションされた地球観測データセットに対する問合せは次の二通りが考えられる

- (1) アノテーションを問合せの入力とし、アノテーションされたデータを返す
- (2) データを問合せの入力とし、入力データに対して行われたアノテーションを返す

が考えられる。本節では 2 番目のアノテーションによって付与された情報を取得するための問合せについて議論する。2 番目の問合せについて一番単純な場合は、特定のデータに関する問い合わせである。一つのデータに対するアノテーションの問合せの入力  $q$  は以下のように表わす

$$q_0 = (ds, s, t, i, v)$$

問合せ  $q_0$  に対しては地球観測データ  $(ds, s, t, i, v)$  に対して付与されたアノテーションを返す。また一定の範囲に分布する地球観測データに対してなされているアノテーションに対する問合せも考えられる。このような問合せを行うには、アノテーションの対象となるデータセットの範囲を指定した時と同様に、各属性に関する選択条件式やドントケアを用いてデータの分布範囲を指定する。範囲を指定した問合せの例を以下に示す。

$$q_1 = (ds, s, t < 1970-01-01, i, *)$$

問合せ  $q_1$  に対しては、問合せの条件を満たす地球観測データのインスタンス全てに対して付与されたアノテーションを返す。

## 4. データの系譜

研究者にとってあるデータセットが有用か否かを判断するために、そのデータの生成工程やデータの出自を参考にすることが考えられるそのような情報は一般的にデータの系譜 (Lineage) と呼ばれる。地球観測データセットの変換は入力となるデータプロダクト、出力となるデータプロダクト及びプロセスの記述

の 3 つの要素で表現することができる。

本節ではまず地球観測データに対して行われた変換を表現するためのデータセットに対する演算を定義を行う。

### 4.1 データセット演算

地球観測データ統合、解析においてデータ集合に対して様々な操作が行われることが考えられる。

本章ではそのような操作をモデル上で表現する、地球観測データセットに対する操作体系について定義していく。地球観測データセットはデータの集合であり、一般的な集合演算、すなわち和集合、差集合、共通集合が定義できる。これらの集合演算に加え、以下に示す演算を定義する。

#### 4.1.1 選択

一般の関係代数演算における選択演算と同様に、選択演算は要素に関して選択条件式を満たすデータのみからなるデータセットの部分集合を求める演算である。データセットや観測項目、スカラー値でない観測値などの属性による選択操作は各属性の要素集合を選択する操作になるが、時間属性や空間属性、スカラー値を持った観測値などに対しては不等式を利用した選択が可能である。選択条件式は以下のように定義される。

[Definition 1]  $X(\in \{DS, S, T, I, V\})$  をデータセット  $D$  の属性、 $Y$  を  $X$  における値の集合、 $\theta \in \{\in, \notin\}$  とするとき、 $X\theta Y$  を  $D$  の選択条件節と言う。また  $V$  の値がスカラー値からなる場合、 $X(\in \{S, T, V\})$  をデータセット  $D$  の属性、 $y$  を定数、 $\theta \in \{<, >, \geq, \leq, =, \neq\}$  を比較演算子とするとき、 $X\theta y$  も  $D$  の選択条件節と言う。

選択条件節のブール式を選択条件式と言う。選択条件式は厳密には以下のように定義される

- (1)  $D$  の選択条件節は  $D$  の選択条件式である。
- (2)  $l$  が  $D$  の選択条件式ならば、 $\neg l$  は  $D$  の選択条件式である
- (3)  $l_1, l_2$  が  $D$  の選択条件式ならば、 $l_1 \wedge l_2$  は  $D$  の選択条件式である
- (4)  $l_1, l_2$  が  $D$  の選択条件式ならば、 $l_1 \vee l_2$  は  $D$  の選択条件式である

$c$  が地球観測データセット  $D$  の選択演算式であるとき、 $D$  の  $c$  に関する選択  $\sigma_c D$  は以下のように定義される。

$$\sigma_c D = \{d | d \in D \wedge c \text{ is true for } v\}$$

#### 4.1.2 集約

集約演算はデータセットの空間解像度又は時間周期、又はその双方の変更を伴う演算である。データセット  $D$  の観測項目  $i$  を持つデータの集合の空間解像度、時間周期をそれぞれ  $\lambda_s(i), \lambda_t(i)$  に変更したデータセットを求める集約演算は以下のように表わされる。

$$\text{opt}_{\lambda_s(i), \lambda_t(i)} D$$

$D$  が空間解像度  $\lambda_{s_0}(i)$ 、時間周期  $\lambda_{t_0}(i)$  を持つ場合、 $\lambda_s(i), \lambda_t(i)$  には以下の様な制約が存在する。

$$\lambda_s(i) \geq \lambda_{s_0}(i), \lambda_t(i) \geq \lambda_{t_0}(i)$$

また、 $opt \in \{max, min, sum, average, count, udf\}$  は具体的に集約を行う方法を表わす。  $max, min, sum, average, count$  はそれぞれ変更後の各空間解像度、時間周期に含まれる  $D$  中のデータ  $d$  の観測値  $v$  の最大値、最小値、和、算術平均、 $d$  の個数を返す演算であり、空間内挿などのその他の演算は  $udf$  と表現する。

#### 4.1.3 統合

統合は二つのデータセットに関して操作を行い、統合条件を満たす新たなデータセットを得る演算である。統合条件は統合によって生成されるデータセットの時間周期、空間解像度、観測項目からなる。

$$D \uplus_{opt, \lambda_s, \lambda_t, i} D'$$

統合演算は二つの入力の各々のデータセットから観測項目  $i$  と同じ観測対象、観測物理量を持つ観測項目を持ったデータを選択し、集約によって時間周期、空間解像度を指定されたものに変換しその操作で得られる 2 つのデータの集合の和集合を取る演算である。統合操作はこれまでに定義された演算の組み合わせ (2 つの入力それぞれに対する選択、集約とその結果の和集合) からなっており、シンタックスシュガーであると言える。また各統合条件において指定がない場合、2 つの入力データセット  $D, D'$  に応じた条件が自動的に設定される。

- $opt$  の指定が無い場合:  $D, D'$  の観測項目属性において観測対象、観測物理量が同じものが存在し、一方にのみ集約方法の要素が存在する場合、その集約方法を用いて集約を行う。双方に集約方法が存在する場合、それぞれの集約方法にて集約を行う。双方とも集約方法が存在しない場合は空集合を結果として返す

- $\lambda_s, \lambda_t, i$  の指定が無い場合:  $D, D'$  の観測項目属性において観測対象、観測物理量が同じものが存在する場合、それらの観測項目を持つデータの空間解像度及び時間間隔の最小公倍数となる値が集約条件に利用される。また観測値の単位は  $D'$  の観測値が持つ単位に揃える。

統合条件の指定が全くない統合を自然統合と呼び、以下のように表現する

$$D \uplus D'$$

自然統合は  $D$  の観測値の単位が  $D'$  のものに変換される可能性があるため、一般に  $D$  と  $D'$  は可換ではない

#### 4.1.4 統合可能性

統合演算を利用して統合可能性という概念が定義できる統合可能なデータセットの定義を以下に述べる

[Definition 2] 2 つの入力データセット  $D, D'$  に関して、統合演算  $D \uplus D'$  の結果が空集合ではない場合、 $D$  は  $D'$  に統合可能であるという

また統合演算および自然結合の定義から、以下の定理が成り立つ。

[Theorem 1] 2 つの入力データセット  $D, D'$  に関して、観測項目属性において観測対象、観測物理量が同じデータが存在しする場合、 $D$  は  $D'$  に統合可能であるという

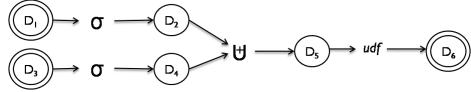


図 1 データ系譜の DAG 表現

統合可能性という概念は、地球観測データ統合解析基盤上で利用者に利用可能なデータセットを提示するのに用いる事が出来る。例えば利用者にとって既知のデータセットに統合可能であるデータセットを提示することによって、観測の空間範囲を補充するデータセットや、時間間隔がより詳細なデータセットの提示が可能である。

#### 4.2 データ系譜モデル

本研究では地球観測データセットの系譜情報を地球観測データセットノードと演算を表わす演算ノードを用いて非巡回有向グラフ (DAG:Directed Acyclic Graph) で表現する。系譜情報の DAG はデータセットノードと演算ノードが交互にあらわれ、それぞれのノードへの矢印はプロセスの出力及び入力を示す。データ系譜情報を表現した DAG の例を図 1 に示す。

図 1 において丸は地球観測データセットを表現しており、実際にシステムが保持しているデータプロダクトは  $D_1, D_3$  からそれぞれ利用するデータを選択し、得られた中間生成データセット  $D_2, D_4$  を統合し  $D_5$  を得、 $udf$  で記述される集約操作を加え  $D_6$  を生成している。

#### 5. 関連研究

地球観測データに依存しない系譜情報のモデリングに関する研究としては [6]~[8] などが存在している。また異なるシステム間におけるデータ系譜情報の相互運用のためのモデルとして、open provenance model [9] のような試みも存在する。

地理情報のモデリングに関しては、XML をベースとした Geospatial Marking Language [10] 等が挙げられる。GML は地理的特性を表現するのに優れているが地球観測データのような科学データを表現するには時間属性や観測項目属性の表現を詳細化する必要がある。地球観測データ相互流通性のためのメタデータ標準としてはアメリカ合衆国連邦地理データ委員会 (FGDC:the Federal Geographic Data Committee) の CSDGM(The Content Standard for Digital Geospatial Metadata) [11] や国際標準機構 (ISO:The International Organization for Standards) によって策定された地理メタデータ標準 [12] などがある。これらの標準においては系譜情報やキーワード情報の記述等に関しては特に制約を定めておらず、システムで利用するには別途詳細に定める必要がある。このような従来のメタデータモデリングは相互流通性を向上するためにその語彙を本研究のアノテーションに利用する語彙として組み込むことが考えられる。

## 6. まとめと今後の課題

本稿では地球観測データに対する柔軟なアノテーションの為のデータモデルと、データ系譜情報の保持についてのモデル化について提案を行った。

今後、データの加工や演算によってアノテーションが受ける影響を考察し、演算後のデータへの自動的なアノテーションの付与について考察を重ねていく。また、実際に提供されている地球観測データに対してモデルを適応し、データ検索やユーザがアノテーションを行うアプリケーションシステムの開発を進めていくとともに、モデルのスケラビリティの検証なども行っていく予定である。

## 謝 辞

本研究は、文部科学省委託業務研究費国家基幹技術「データ統合・解析システム」の支援を受けており、ここに記して謝意を表します。

## 文 献

- [1] Netcdf(network common data form).  
<http://www.unidata.ucar.edu/software/netcdf/>.
- [2] Grid analysis and display system (GrADS).  
<http://www.iges.org/grads/>.
- [3] Science and technology facilities council.  
<http://www.scitech.ac.uk/>.
- [4] RP Tyer, PA Couch, K Kleese van Dam, IT Todorov, RP Bruin, TOH White, AM Walker, KF Austen, MT Dove, and MO Blanchard. Automatic metadata capture and grid computing. In *Proceeding of the UK e-Science All Hands Meeting*, pp. 381–384, 2006.
- [5] 高橋慧, 絹谷弘子, 吉川正俊. オントロジを利用したメタデータ構築に基づく地球観測データ統合検索フレームワークの研究. データベースと Web 情報システムに関するシンポジウム (DBWeb2007), 2007.
- [6] Yingwei Cui and Jennifer Widom. Lineage tracing for general data warehouse transformations. *VLDB J.*, Vol. 12, No. 1, pp. 41–58, 2003.
- [7] Peter Buneman, Adriane Chapman, and James Cheney. Provenance management in curated databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 539–550, 2006.
- [8] Rajendra Bose. A conceptual framework for composing and managing scientific data lineage. In *SSDBM*, pp. 15–19, 2002.
- [9] Open provenance model. <http://openprovenance.org/>.
- [10] Geograpy markup language (gml).  
<http://www.opengeospatial.org/standards/gml>.
- [11] Federal Geographic Data Committee. Content standard for digital geospatial metadata. fgdc-std-001-1998,, June 1998.
- [12] International Organization for Standardization. Iso 19115:2003, geographic information metadata.