社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

地球科学データアノテーションシステムの構築

高橋　慧†　立床　雅司†　絹谷　弘子††　吉川　正俊†

† 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
†† 東京大学 地球観測データ統融合連携研究機構　〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{atakahashi,m.tatedoko}@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，地球温暖化や異常気象などの問題が顕在化し，地球科学関連研究の重要性が一層増してきている．今後，地球環境への理解をさらに深め，地球環境問題の解決や災害対策につなげるには，気候や水循環，農業などの多様な分野で観測が行われているデータを共有し，分野を跨いだデータの統合解析を行うための基盤が必要である．しかし，地球科学データの多くは機関やプロジェクトによってアドホックな形態で保存されている場合が多く，研究者が専門外のデータを検索，利用するためには個々のデータに対する十分な理解が必要となる．そのため，各種データの利用方法などが把握できるよう，メタデータを充実させる必要がある．本研究では我々が開発している地球観測データ統融合システムにおいて登録されているデータとメタデータを関連付けるアノテーションシステムについて議論する．我々が構築を目指すシステムにおいては，地球科学データに対して柔軟な情報のアノテーションを提供したり，アノテーションされた情報を自動的に関連するデータプロダクトに伝播させる機構を設けることにより，データ提供者のメタデータ構築作業を軽減したり，データ利用者による自由な情報のアノテーションおよび共有を実現する．
キーワード　メタデータ，アノテーション，データ系譜，科学 DB

Development of an Annotation System for Earth Science Data

Akira TAKAHASHI†, Masashi TATEDOKO†, Hiroko KINUTANI††, and Masatoshi YOSHIKAWA†

† Graduate School of Informatics, Kyoto University　Yoshida-honnmachi, Sakyo-ku, Kyoto, 606–8501 Japan
†† Earth Observation Data Integration and Fusion Research Initiative (EDITORIA), the University of Tokyo
Komaba 4–6–1, Meguro-ku, Tokyo, 153–8505 Japan
E-mail: †{atakahashi,m.tatedoko}@db.soc.i.kyoto-u.ac.jp, ††kinutani@tkl.iis.u-tokyo.ac.jp,
†††yoshikawa@i.kyoto-u.ac.jp

Abstract　Earth science researches have become increasingly important, as environmental issues such as global warming and climatic aberration has gained people's interests. Now, to promote further understanding of the earth environment and solve earth environmental problems, analysis and integration of data from diverse research domain, such as climatology, hydrology, agriculture, is necessary. However, most data researchers find difficulties in discovering and interpreting useful data from heterogeneous datasets. In order to support research activities, enhancement of metadata for data discovering and profound understanding of data is a crucial issue. In this paper we discuss about the data annotation system we are building over our earth observational data integration and analysis system. We aim to provide flexible annotation methods for earth science data and enable annotations to propagate among related data products in order to reduce data management costs, and allow users to share information annotated to earth science data.
Key words　Metadata, Annotation, Data Lineage, Scientific database

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く，また地球科学データは分野やプロジェクトに固有な，アドホックな形式で保存されている場合が多く，実際に利用する為にはデータについての十分な理解が要求される．そこで，データの検索や内容理解を支援するために，地球科学データのメタデータを充実させる必要がある．そのためには，データに対してユーザが自在にアノテーションを行い，ユーザ間で共有するシステムを構築することが有用であると考えられる．ユーザがアノテーションを行うシステムの成功例としては，Web2.0 と呼ばれるユーザがコンテンツを投稿し，コンテンツにタグやコメント等をアノテーションするやゲノム解析において DNA に遺伝子情報を付与するシステムなどがあげられる．地球観測データに対しても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．分野にまたがった地球観測データに対するアノテーションシステムはまだ確立されていないのが現状である．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をおこなうシステムの実現にむけて議論する．

本稿の構成は以下の通りである．まず 2. で関連研究について紹介し，**??**で地球科学データアノテーションシステムに必要な機能について議論する．3. で本研究で用いるデータモデリングについて説明し，4. で構築を目指しているシステムの概要を提示,．最後に 5. でまとめを行う．

2. 関連研究

本節では，メタデータのアノテーションを行うシステムの現状について紹介する．各種データの理解促進やクラスタリング，意味付けのためにアノテーションを行うシステムは多数提案されている．Web 上に散らばるコンテンツに対するアノテーションを行う研究としては，Annotea [1] や [2], [3] などが知られている．また，delicious(注1)やはてなブックマーク(注2)等に代表されるソーシャルブックマークサービスや Youtube(注3)，Flikr(注4)などのマルチメディアコンテンツ共有サービスにおいては，利用者によってタグやコメントがアノテーションされることにより，多種多様なコンテンツの分類や評価が可能になっている．

バイオインフォマティクスの分野では統制語彙である Gene Ontology [4] を用いて, 遺伝子の塩基配列や文献などへの機能アノテーションが行われている．また [5], [6] などのシステムによってアノテーションの共有が行われている．

地球科学データに関するアノテーションとしては，相互流通のためのメタデータ仕様がいくつか提案されているアメリカ合衆国連邦地理データ委員会 (FGDC:the Federeal Geographic Data Committee) は地理情報標準として CSDGM(The Content Standard for Digital Geospatial Metadata) [7] を策定しており，国際標準機構（ISO:The International Organization for Standards）では地理メタデータ標準 ISO19115:2003 [8] が策定されている．これらの標準においてはデータの系譜情報や観測項目情報，キーワード情報の記述等に関しては特に制約が定めて無いため，システムで利用するにはメタデータを拡張する必要がある．これらのメタデータ標準に準拠したメタデータ構築システムは, MetaLite(注5), CorpsMet(注6), NOAA ArcView Extension(注7), CatMDEdit(注8) など多数存在する．またメタデータ標準にしたがったメタデータの共有を支援するツールとして GeoNetwork(注9)が存在する．GeoNetwork は国際連合食料援助機関 (FAO) が開発したオープンソースのメタデータカタログシステムである．GeoNetwork サーバを GeoNetwork のノードとして登録することで他の GeoNetwork ノードとメタデータを交換することができ，クリアリングハウスの構築が容易にできる．

この様にメタデータ標準に従ったメタデータを構築することで既存のツールを利用できるという利点は大きい．しかしこういったメタデータ標準やそれを利用したシステムは主に地球科学データカタログの構築が目的であり，メタデータ標準の枠組みの中では例えば複数データセットに対してアノテーションを行ったり，多数のユーザによるアノテーションなどは実現が難しいのが現状である．またメタデータ標準に従うことで，今後標準の枠組みに収まらない情報の付与に対応できない可能性も考えられる．

3. 地球観測データに対するアノテーション

本節では本研究で用いるアノテーションのモデルについて説明を行う．

(注1)：http://delicious.com/
(注2)：http://b.hatena.ne.jp/
(注3)：http://www.youtube.com/
(注4)：http://www.flickr.com/
(注5)：http://nsif.org.za/Metadata/metadataall.htm
(注6)：http://corpsgeo1.usace.army.mil/
(注7)：http://www.csc.noaa.gov/metadata/download.html
(注8)：http://catmdedit.sourceforge.net/
(注9)：http://geonetwork-opensource.org/

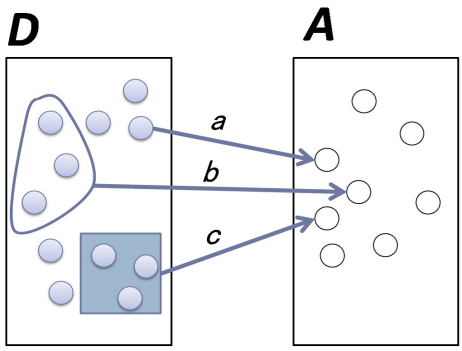

図 1　アノテーションと地球科学データ

3.1　データアノテーション

Web 上のコンテンツにおけるアノテーションは，w3.org における議論によると以下のようなものであるとされている [9]

> Any object that is associated with another object by some relationship

URI によるオブジェクトの一意識別が可能な場合はこの様な位置付けでも問題無いと考えられるが，地球科学データに対するアノテーションとしては，例えば同じ地域で観測された別のプロジェクトのデータや，ある一定の期間に観測されたデータなど，多様な粒度を持ったアノテーションが考えられる．そこで，我々は図 1 に示されるように，アノテーションを地球科学データオブジェクトの集合とアノテーションオブジェクトの関係としてとらえる．

以下，地球科学データオブジェクトとアノテーションオブジェクトについて説明を行う．

3.1.1　地球科学データオブジェクト

我々が扱う地球科学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 [10],GrADS [11] 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．これらの多様なデータに対して一元的にアノテーションする要求に応える場合，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．

我々が扱う地球科学データには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が最低限紐付けられているものとし，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**??**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決める．

（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):**
観測された値がどういう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の温度の事を指すが，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 3 つの観点 (aspects) 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．

これらの項目を表現するために，先行研究 [12] で示したように，NASA のジェット推進研究所によって整備されており，すでにいくつかの地球科学プロジェクトでの利用実績 [13], [?], [14] がある SWEET [15] オントロジを利用し，RDF で表現する．観測項目の各観点とプロパティ及びそのドメインとなるクラスが含まれるオントロジとの対応を表 2 に示す．RDF を用いて一定期間，地域における最高気温を表す観測項目を記述した例を図 2 に示す．

（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．

図 3 に地球観測データのインスタンスを示す．図 3 のインス

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
:	:	:	:	:	:	:	:	:	:	:

表 1　Example of earth observation data: WMO Resolution 40

観測項目の観点	Property	domain ontology
観測対象	phenomena:hasAssociatedPhenomena substance:hasAssociatedSubstance	SWEET phenomena SWEET substance など
観測物理量	property:hasAssociatedQuantity	SWEET property
集約方法	ex:aggregatedBy	SWEET numerics

表 2　SWEET オントロジを用いた観測項目の記述

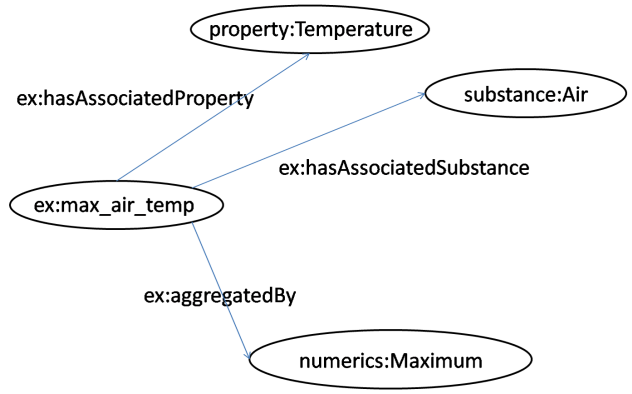

図 2　観測項目「最高気温」の RDF 表現

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

図 3　Examples of data

タンスは表 1 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．なお地点や時刻，観測項目情報などの表記は簡易的なものであり実際の実装とは異なる．

3.1.2　アノテーションデータオブジェクトとアノテーション

本研究ではアノテーションデータオブジェクトを以下の 3 つ組で表現する．

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

u はアノテーションを行った利用者の識別子を表し，t はアノテーションが有効な期間 (valid-time) を表す．c はアノテーションの内容を表す．アノテーションデータオブジェクト a による地球科学データオブジェクト集合 D のアノテーション A を $A = (a, D)$ で表す．アノテーションを行う場合には具体的な集合 D を指定することになるが，先に挙げた例のように，同じ地域で観測された別のプロジェクトのデータや，ある一定の期間に観測されたデータに対してアノテーションを行いたい場合，実際にユーザが指定した条件を満たすデータからなる集合を構成する必要がある．また，システムに新しくデータが投入された場合はユーザの意図に従って，D を新しく構成し直す必要がある．この様な手間を省くために，本研究では地球科学データオブジェクトの集合にアノテーションを行うだけではなく，地球科学データオブジェクトの満たす条件に対してアノテーションを行う事を許す．地球科学データオブジェクトの満たす条件は各属性に対して以下の条件を指定する事で表現する，

（1）　任意の値を許す
（2）　選択条件式を満たす値を許す
（3）　特定の値のみを許す

例えばデータセット識別子，時間属性，観測項目属性に特定の値のみを許し，空間属性および観測値に任意の値を許すといった条件を満たすデータセットに対するアノテーションは以下の

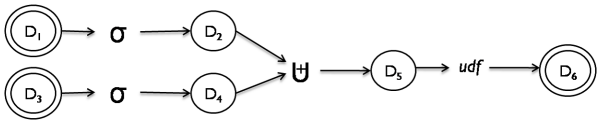

図 4　データ系譜の DAG 表現

ように記述する．

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

またデータセット識別子，空間属性，観測項目属性に特定の値のみを許し，観測値に任意の値，時間属性にある値 X より大きい値を許すといった条件を満たすデータセットに対するアノテーションは以下のように記述する．

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

3.2　データ系譜モデル

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，データの生成行程や，各行程における処理アルゴリズムの情報，処理の責任者の情報等を合わせてデータの系譜 (Lineage) と呼ぶ．データの系譜を保存しておくことで，データの作成に寄与した研究者の責任が明確になる．また系譜上のあるデータにアノテーションした場合，同じ系譜上に存在するデータに対してアノテーションを伝播させ，各データにアノテーションを行う手間を省くことができる．

本研究では地球科学データの系譜情報を地球科学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 4 に示す．各演算については [16] で定義されたものを利用している．

図 4 において丸は地球科学データオブジェクトの集合を表現している．実際にシステムが保持しているデータプロダクトはトは 2 重丸，データを生成する際に生じた中間生成データセットは 1 重丸で表現している．図 4 に示したプロセスではデータセット 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．

この DAG 表現では各データオブジェクト集合間の関係のみが記述されており，一般的なデータ系譜情報の表層的な表現にとどまってしまっている．そのため，各データオブジェクト集合に対して適切なアノテーションを行う必要がある．本研究では以下の情報を DAG 中の各オブジェクト集合に対してアノテーションすることで，データ系譜情報を表現する．

（1） オブジェクトの生成に用いられた具体的な処理，アルゴリズムなどの情報

（2） オブジェクトの生成の責任者

4.　地球科学データアノテーションシステム

本節では前節までに提案したデータモデルを利用した地球科学データアノテーションシステムについて紹介を行う．我々が構築を目指すシステムの概要図を図 5 に示す．図 5 において一番下の層に実際の地球科学データのストレージが存在しており，地球科学データは任意のモデル，ファイル形式で保存されている．利用者やクライアントアプリケーションは保存に用いられているデータモデルと，地球科学データオブジェクトモデルとのマッピングを行うデータアクセスメディエータを通じてデータにアクセスを行う．これによりクライアント側は実際のデータモデルを意識せずにデータにアクセスできる．

またアノテーションデータオブジェクトはアノテーションの対象となる条件式や集合の記述と組でメタデータデータベースに保存され，メタデータの検索はそのコンテンツに対する検索に加え，地球科学データオブジェクトを入力とし，該当データオブジェクトに対してなされたアノテーションを返すといった検索手法を実現する．

我々が扱う地球科学データプロダクトの一部は有償であり，データのアクセス制御は厳密に行わなければならない．また利用者が投入するメタデータも作成者の意図に沿った範囲でのみ公開される必要がある．

従って，地球科学データおよびメタデータにアクセスするには実際にアクセス権限があるかどうかのアクセス権限判定モジュールを経て行う必要がある．データ検索アプリケーションやデータ閲覧アプリケーションなどのクライアントソフトウェアはこのアクセス権限判定モジュールの上に構築される．

4.1　利用シナリオ

我々が構築を目指しているシステムにおいては，メタデータの多くは利用者に入力してもらうモデルになっている．しかしメタデータの入力にはコストを要し，実際にメタデータを登録してもらうにはそのためのインセンティヴが必要である．我々はデータ提供者とデータ利用者に対してそれぞれ有益なツールを用意することによって．インセンティヴとすることを試みようと考えている．

データ提供者に対しては，入力したデータ利用のためのメタデータを元に相互流通性のためのメタデータや，メタデータを紹介するドキュメントを自動生成するツールを提供する事が考えられる．また，

データ利用者に対してはソーシャルブックマークサービスにおけるタグ付け，コメント付与といった機能を提供し，自分の研究のためのデータの整理ツールとしてアノテーションシステムを提示することが考えられる．

5.　まとめと今後の課題

本稿では地球科学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構築を目指す地球科学データアノテーションシステムについて説明を行った．現在システムの実装を進めている段階であり，今後実際に科学者の方に利用していただきシステムの有効性を定性的に考察するとともに，システムのスケーラビリティについても検証を行っていく予定である．また実際に投入されたメタデータの活用についても今後議論していく．

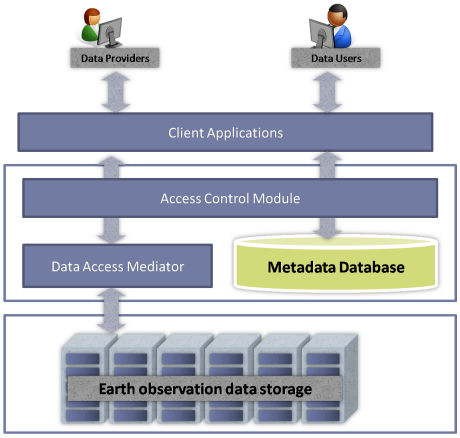

図 5　地球観測データ管理システム概要

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] José Kahan, Marja-Riitta Koivunen, Eric Prud'hommeaux, and Ralph R. Swick. Annotea: an open rdf infrastructure for shared web annotations. *Computer Networks*, Vol. 39, No. 5, pp. 589–608, 2002.

[2] Daniel LaLiberte and Alan Braverman. A protocol for scalable group and public annotations. *Computer Networks and ISDN Systems*, Vol. 27, No. 6, pp. 911–918, 1995.

[3] Matthew A. Schickler, Murray S. Mazer, and Charles 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.

[4] Gene ontology. http://www.geneontology.org/.

[5] biodas.org. http://biodas.org/.

[6] Robin D. Dowell, Rodney M. Jokerst, Allen Day, Sean R. Eddy, and Lincoln Stein. The distributed annotation system. *BMC Bioinformatics*, Vol. 2, p. 7, 2001.

[7] Federal Geographic Data Committee. Content standard for digital geospatial metadata. fgdc-std-001-1998,, June 1998.

[8] International Organization for Standardization. Iso 19115:2003, geographic information metadata.

[9] Collaboration, knowledge representation and automatability, w3c. http://www.w3.org/Collaboration/.

[10] Netcdf(network common data form). http://www.unidata.ucar.edu/software/netcdf/.

[11] Grid analysis and display system (GrADS). http://www.iges.org/grads/.

[12] 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.

[13] The geoscienes network (geon) project. http://www.geongrid.org/.

[14] Virtual solar-terrestrial observatory (vsto). http://vsto.hao.ucar.edu/.

[15] Semantic web for earth and environmental terminology (sweet). http://sweet.jpl.nasa.gov/ontology/.

[16] 高橋慧, 絹谷弘子, 吉川正俊. 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. iDB フォーラム 2008 情報処理学会研究報告「データベースシステム」, Vol. Vol.2008 No.88, pp. pp.25–30, 2008.