地球観測データ統合利用支援のためのデータ品質のモデリング

高橋　慧† 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
†† 東京大学生産技術研究所 〒 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，様々な分野において関連機関における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を提供が可能となる．研究者の要求を満たすデータを判断するためにはデータの質の可視化は必須である．本研究では解析，統合利用に有用なデータの質を表現し，多様な地球観測データに柔軟な適応が可能なメタデータモデルを提案し，その利用手法と有用性を考察する．

キーワード　データモデル，メタデータ管理，科学 DB

Modeling the Quality of Data for Integrated Use of Earth Observational Data

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

† Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Institute of Industrial Science, The University of Tokyo
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　As the technology of earth observation have been developed in the past years, large amount of Earth observation data has been acquired and stored by institutes among various fields of researches. Founding an infrastructure to integrate and analyze such Earth observation data to produce valuable data products has gained interests. In such systems, visualizing the quality of data is necessary to help scientists identify data which meet their requirement of analysis and experiments. This article describes what kind of information is available for scientists to measure the quality of a data, shows a representation model of such quality related metadata, and its application to actual distributed data.

Key words　Data Model, Metadata, Scientific Databases

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等における測定機器の観測精度の上昇，地球シミュレータなどの大規模計算機インフラの整備などにより，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多角的に統合，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を提供が可能となると考えられている．

そのような基盤を構築，運用する際に，あるデータを研究者が解析，実験の要求を満たし，実際に利用するか否かを判断するためには，そのデータの品質にまつわる情報が提示される必要がある．しかし既存のデータの品質情報の表現方法，メタデータ形式は当該分野によって異なり，研究者は専門外の分野に関するデータの品質について十分に理解することができない．そこで分野の多様性を吸収し，データの品質を統一的に表現するためのメタデータが必要となってくる．本研究では解析，統合利用に有用であると考えるデータの品質にまつわる情報を表現し，多様な地球観測データに柔軟な適応が可能なメタデータモデルを提案し，その利用手法と有用性を考察する．

本稿の構成は以下の通りである．まず 2. で地球観測データの品質について考察を行い，関連するメタデータ標準の紹介を行う．3. で本研究におけるメタデータのモデリング手法の説明を

示し，4. でメタデータモデルの XML による実装例を示し，有用性や利用方法について考察する．5. で今後考察すべき事項を示し，最後に 6. でまとめを行う．

2. 研究の背景

2.1 地球観測データの質

地球観測データは一般に，地理空間情報の一種であり，観測機器の出力データに地理空間情報や観測時刻情報が付随するデータであり，付随データや観測条件の説明などのためにメはタデータが利用される．地理空間，観測時刻といったメタデータは主にデータの発見用途に用いられ，科学者が求める空間範囲や期間に存在するメタデータを検索するのに用いられる．それに対してデータの品質は発見したデータが実際に利用できるのかを判断するメタデータとして表現が可能であるといえる．データ品質は多様な属性からなる．地球観測データの品質としては例として以下のようなものがあげられる．

- **データの精度:** 観測されたデータの有効数字，データの時間軸，空間軸における解像度，付随する地理情報，時間情報の正確性等．
- **データの起源:** データの観測に使われた機器情報やその環境情報，観測責任者, 観測生データからどのような加工が行われてきたか，等．
- **データの状態:** データの量，完全性, 論理一貫性等
- **データの評判:** データの利用頻度，論文への引用頻度，データの使いやすさ等．

データの品質という概念は非常に広範にわたり，定量的な評価が困難である属性も多い．

2.2 メタデータ標準におけるデータの質

地理情報標準におけるメタデータ標準としては，1994 年にアメリカ合衆国連邦地理データ委員会 (FGDC: the Federal Geographic Data Commitee) が策定した CSDGM(Content Standard for Digital Geospatial Metadata) [1] がある．CSDGM ではデータの品質について以下の五つの次元で表現している．

- **属性正確度 (Attribute accuracy):** 属性が地理的な性質を正しく表現しているか
- **論理一貫性 (Logical consistency):** 要求される仕様をデータが満たしているか
- **完全性 (Completeness):** 必要なデータを網羅しているかどうか
- **位置正確度 (Positional accuracy):** データセットにおける位置情報と実際の位置との差異
- **データ系譜 (Lineage):** データが取得，生成，加工された工程の詳細な記録

CSDGM は 1998 年に改訂され，リモートセンシングによって得られたデータセットにおける雲の割合がデータ品質を表す次元として追加された．2003 年には FGDC CSDGM を元に国際標準機構（ISO:The International Organization for Standards）の TC211（地理情報専門委員会）によって地理メタデータ標準である ISO 19115 [2] が策定された．ISO19115 のデータの質

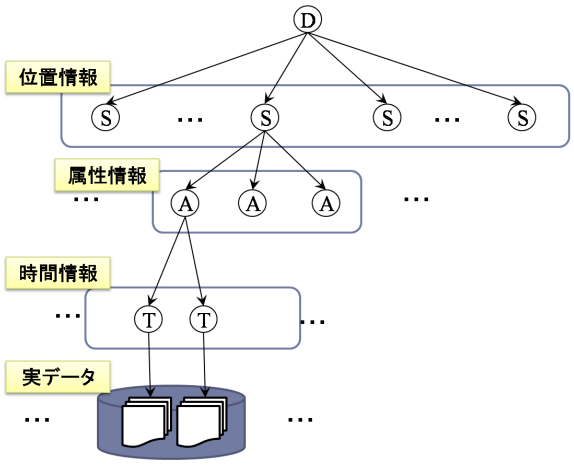

図 1　データ分類木構成の例
Fig. 1　Example of Data tree

への考え方は CSDGM とほぼ同じ方針であり，上記の 5 項目に時間正確度 (temporal accuracy) が加えられ, 属性正確度が分類正確度 (thematic accuracy) に変更されている．日本では ISO19115 等を基に，地理情報標準第 2 版 (JSGI2.0: Japanese Standards for Geographic Information 2.0) [3] 地図情報標準プロファイル (JPGIS) Ver.1.0 [4] などの地理情報標準が策定されている．

メタデータ標準はファイル保存形式や言語情報などの，データ利用のための基本的な属性に関しては相互流通性を確保するために有用であるといえる．しかし，データの品質の記述に関してはまだまだ議論が活発であり，メタデータ標準で策定，利用されるデータの品質情報は多様な属性からなるデータ品質の一部のみにとどまっており，データ品質のさらなるモデリング手法の検討は必要である [5].

また ISO19115 で必須項目とされている属性 (データ集合の要約やメタデータ問合せ先) 等は細かい粒度でメタデータを管理する場合は重複が生じ冗長であり，実際に運用する際にはデータの構造を考慮した柔軟なメタデータ付随戦略が必要であると考えられる．そこで本研究では多様なデータ粒度に対してデータの品質を付随するメタデータモデルの構築を目指す．

3. メタデータモデリングのアプローチ

一般に，あるデータセット内の地球観測データは位置属性，時間属性や観測項目属性情報によって分類することができ，属性を一定の順序で評価しデータを分類する分類木を考えることができる．図 1 はデータセット内のデータを位置属性，観測項目属性，時間属性の順で構成される評価軸を表す分類木を構築した例である．

この分類木を用いてデータの構造を表現するメタデータを構築することができる．これを構造メタデータとして定義する．構造メタデータは図 2 に示すデータ粒度オブジェクト (Data-Granule) をノードにもつ木構造で表現される．粒度オブジェクトである各クラスの説明を以下に記す．

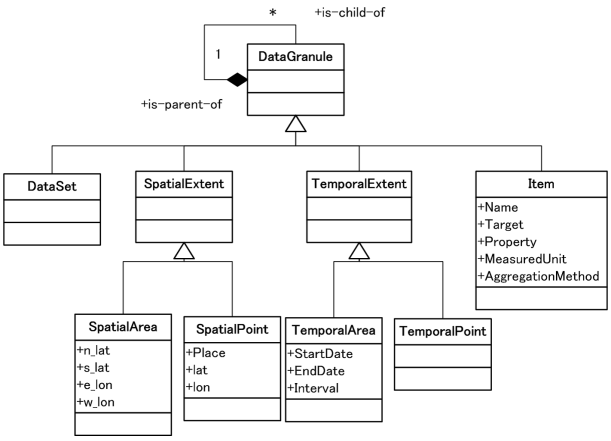

図 2　データ粒度オブジェクト
Fig. 2　Data Granule Object

- **データセット (DataSet):** 観測機関やプロジェクト内などのデータのまとまりを示す
- **位置情報 (SpatialExtent):** 観測が行われた地点，領域.
- **時間情報 (TemporalExtent):** 観測が行われた期間を表す
- **観測項目情報 (Item)** 観測対象となる物質，現象やその物理量，観測に使われた機器などを表す.

構造メタデータの木構造において同じ深さに存在する粒度オブジェクトは同じクラスである.

構成される構造メタデータは実際のデータの保存形式によらない概念的な構造を表している．したがって構造メタデータで示された階層構造と実データの保存形式に物理的な階層構造とがそのまま対応していない場合がある．このためメタデータと実データを結びつけるには，実データを，観測項目，位置情報，時間情報を指定して取得するサービスで覆い，メタデータと実データの結びつきはそのサービスを介して行う.

3.1　構造メタデータと品質メタデータの対応

地球観測データに紐付けられる品質情報には，下位の粒度で共通する情報が付随する場合がある．例えば観測機器の精度情報などは同じ観測項目を持つデータに共通する．データの起源情報などはデータセット全体で共通する場合も考えられる．また Tobler の地理学第一法則 [6] に言われるように，近隣のデータほど近似した性質をもち，一定のデータの粒度で品質情報を共有することがあると考えられる．その際品質情報保存の冗長性を減らすために，品質情報は構造メタデータモデルにおいて下位階層へ継承すると考え，品質情報を粒度に応じたノードに付随させる.

半構造データに品質情報を付加するモデルとしては D^2Q モデル [7] 等があるが，本研究では同様に半構造モデルにデータを付与した後，さらに継承関係を考えることによってメタデータのサイズ最適化を行える

以下に例として，図 3 で示される構造メタデータを持つデータセットの品質情報を保存する品質メタデータの構築手順を記述する.

（1）　品質情報の対象となるデータの範囲を表す構造メタデータの葉に，品質情報ノードを付随する．品質情報の対象が，葉が表すデータの粒度より小さい範囲のデータである場合は，その範囲を包含した葉ノードの子にさらに粒度オブジェクトを追加し，その粒度オブジェクトに品質情報ノードを付随させる.

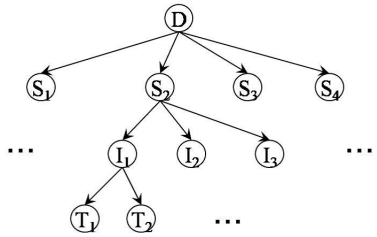

図 3　構造メタデータの例
Fig. 3　Example of a structure metadata

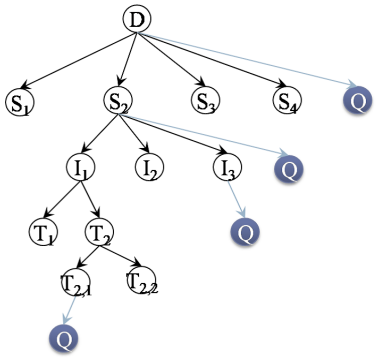

図 4　品質情報を追加した構造メタデータ
Fig. 4　data quality object

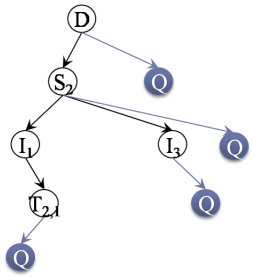

図 5　品質メタデータの例
Fig. 5　Example of a quality metadata

（2）　子孫ノードすべてに同じ品質情報オブジェクトが付随している粒度オブジェクトノードは，子孫ノードから品質情報ノードをはずしそのノードに付随させる．この操作を，再び行えなくなるまで繰り返す (図 4).

（3）　品質情報ノードが付随する粒度オブジェクトノードに，そのノードの兄弟内での順番を表す情報を付随する.

（4）　品質情報ノードと根との経路上のノード以外のノードを削除する.

（5）　品質情報ノードと根との経路上の粒度オブジェクトで，同じクラスが連続しており，かつ親ノードが他の品質情報ノードと根との経路上にない場合は親ノードを削除し，親ノードに張られていた枝を子ノードに張る．(図 5).

ここで，品質情報ノードは図 6 に示すデータ品質オブジェクト (DataQuality) である．データ品質オブジェクトには完全性，論理一貫性に加え，以下の二つの次元を表すクラスを加えた．

データ系譜 (Lineage)　メタデータの対象データプロダクトがどのようなプロセスを経て生成されたか．多くのメタデータ標準ではこの次元は簡単な文章による表現にとどまっていたが，データ統合解析基盤上では先祖データの URI や，加工プロセスを WS-BPEL 等のプロセス記述スキーマを利用した表現 (への URI) など具体的な値を利用する．

データ利用状況 (Usage)　対象データプロダクトの利用状況．これまでは地球観測データの利用状況は論文などに引用されて初めて利用実態が確認できる場合が多かった．データ統合解析基盤上ではそのような引用情報に加え，データの統合利用情報やダウンロード情報などのメタデータも取得が可能である．データの利用実態の多さは，コミュニティにおけるデータの有用性の評価として見ることができ，このような情報の記載はデータ観測者のデータ提供へのモチベーションにもなりうると考えられる．

なお利用可能な品質情報クラスは図 6 に存在するクラスに限らず，ISO19115 において表現可能な位置正確度や時間正確度等も利用可能であり，今後さらに有用な品質情報について検証し追加していく予定である．

4. メタデータモデルの適用

本研究で設計したメタデータモデルを中央農業総合研究センター(注1) が公開している FieldServer (注2)観測データからなるデータセットに適用し検証を行う．FieldServer 観測データは観測点によって観測周期や観測項目が異なるため，構造メタデータの評価軸は，観測地点, 観測時期, 観測項目の順に取った．

今回提案した構造メタデータモデル及び品質メタデータモデルは木構造になっているため，XML による実装が可能である．図 7 にデータセットの構造メタデータインスタンスの一部, 図 8 に品質メタデータインスタンスの一部を示す．

FieldServer データセットにおいてはこれまでデータの蓄積は行われていたが，データセット全体で観測日程や測定の漏れを表現するメタデータが存在していなかったが，今回提案したメタデータモデルを適用することによりそのような情報を管理できるようになった．細かい粒度によるメタデータ管理により，今後はデータのクオリティコントロールシステム [8] 等によって得られた情報なども表現可能であり，結果をユーザやサービス間で共有する等の利用が考えられる

5. メタデータ管理と多様性

メタデータのモデリングは多様な次元をもつデータを実体をいかに効率的に利用しやすい形に落とし込むかを考慮する工程である．本稿では構造メタデータの評価軸にそってデータの概念的な階層構造を構築することで，管理や探索にかかるコストを軽減できる試みを示した．しかしある一つの評価軸を与えると，付加した品質情報のセマンティクスが失われる場合がある．例えば，観測地点，観測時刻，観測項目の順からなる評価軸を用いて構造メタデータを構築した場合，異なる地点の同じ観測項目にある情報を付与する場合は各観測地点以下の観測項目を表すノードに情報を付与すればよいのだが，この場合データセット以下の観測項目に情報を付与したというセマンティクスが失われてしまい，メタデータをマイニングすることでしかその情報を見いだせない．これを解決するには別の評価軸，例えば観測項目，観測地点，観測時刻の順からなる評価軸をもったメタデータインスタンスを新たに構築すればよいのであるが，複数のメタデータ間の整合性を保つコストが増加してしまう．

メタデータ保持コストを増大させて多様な側面からデータを表現するのか，それともメタデータ標準によりメタデータ管理コストを下げ，メタデータやデータのマイニングによって知見を得るのかのトレードオフであり，メタデータ標準などは後者の立場をとっていると言える，しかし地球観測データの多角的な解析による知見のセマンティクスを表現するには複数の評価

```
<?xml version="1.0" encoding="utf-8"?>
<dataset name="fieldserver" >
  <spatialExtent place="fs01"
                 lat="36.51"
                 lone="138.47">
    <temporalExtent startdate="2007/01/31; 17:34:01"
                    enddate="2007/04/26; 23:58:24"
                    interval="2min">
      <item name="Air-temp."
            target="air"
            property="temperature"
            measuredunit="degreeC">
      <item>
      </item name="Humid"
             target="air"
             property="relativehumidity"
             measuredunit="percent">
      </item>
      <item
         ...
      </temporalExtent>
  </spatialExtent>
  <spatialExtent place="fs02"
                 lat="36.51"
                 lon="138.47">
         ...
  </spatialExtent>
</dataset>
```

図 7　構造メタデータインスタンス
Fig. 7　An instance of Structual Metadata

(注1)：http://narc.naro.affrc.go.jp/
(注2)：FieldServer. http://model.job.affrc.go.jp/FieldServer/

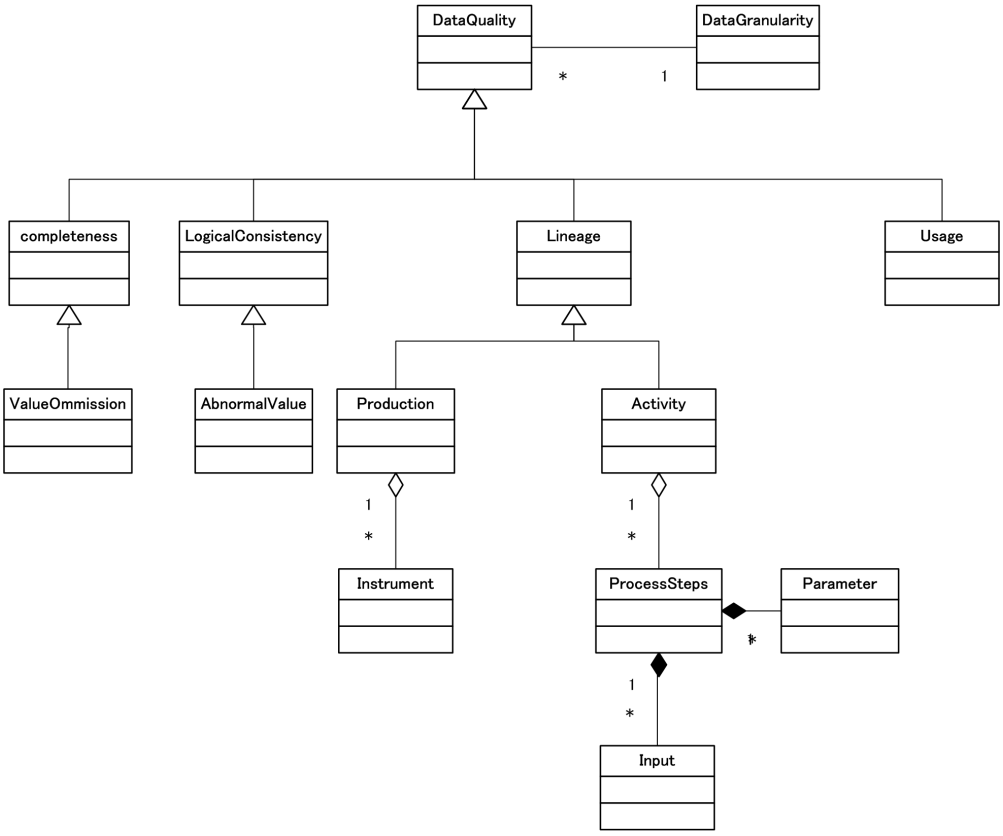

図 6　データ品質オブジェクト
Fig. 6　data quality object

```
<?xml version="1.0" encoding="utf-8"?>
<dataset name="fieldserver">
  <spatialExtent place="fs01"
                 lat="36.51"
                 lone="138.47">
    <temporalExtent startdate=" 2007/04/12; 03:05:40"
                    enddate="2007/04/13; 20:04:38">
        <dataquality>
          <valueomission />
        </dataquality>
    </temporalExtent>
     ...
</dataset>
```

図 8　品質メタデータインスタンス
Fig. 8　An instance of Quality Metadata

軸からなるメタデータ群の構築が必要であり，今後その有用性について検証する必要がある．

6. まとめと今後の課題

本稿では，多様な分野で観測される地球観測データを統合，解析する基盤における，データの品質にまつわる情報を表現するためのメタデータスキーマを提案した．今後はさらに多様なデータセットを投入し，モデルの妥当性の検証およびモデルの改善を行う予定である．また本稿で構築したメタデータモデルに対する品質情報の問合せについて考察を行い，問合せ利用により表現可能なインタフェースの構築，利用状況や系譜情報メタデータの自動生成を行う枠組みを構築，運用し実際に科学者に利用していただき，得られた知見をフィードバックして行こうと考えている．

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] Federal Geographic Data Committee: “Content standard for digital geospatial metadata. fgdc-std-001-1998,” (1998).
[2] International Organization for Standardization: “Iso 19115:2003, geographic information metadata”.
[3] 国土交通省国土地理院：“地理情報標準第 2 版” (1999).
[4] 国土交通省国土地理院：“地理情報標準プロファイル (jpgis) ver.1.0” (2005).
[5] M. F. Goodchild: “Towards user-centric description of data quality.”, Spatial Data Quality 2007 (International Symposium on Spatial DataQuality), Enschede, Netherlands, June 13-15. (2007).
[6] W. R. Tobler: “A computer movie simulating urban growth in the Detroit region”, Vol. 46(2), pp. 234–240 (1970).
[7] M. Scannapieco, A. Virgillito, C. Marchetti, M. Mecella and R. Baldoni: “The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems”, Information Systems 29, Vol. 7, pp. 551–582 (2004).
[8] 生駒, 玉川, 小池, 喜連川：“大規模地球環境観測データを対象と

したデータクオリティコントロールシステムの構築とその有効性の検討”, DBSJ Letters, **4**, 1, pp. 57–60 (2005).