# 地球惑星科学メタデータのデータベースとの連携

寺薗　淳也
Terazono, Junya

会津大学
terazono@u-aizu.ac.jp

# メタデータとは、結局

- **メタデータとは、データに付随するものである。**
  - データのより大本の部分をたどるから、メタ(抽象的な上位)データである。
  - それ自身もデータの形をとっているから、基本的にはデータと見なされる。
- **データに付随するものであるにもかかわらず、データとは別に存在するものである。**
  - 例えば、本をデータとするならば、著作者や表題のような、データを抽象化するものであるがそのデータを表すのに役立つ概念を指している。


# データ構造とは、そもそも

- データ構造は、そもそもデータ本体とメタデータの部分に分かれる。
  - あらゆるデータは、本体の部分と、メタデータとしての2つの部分に分かれている。
  - たいていのデータは、本体の部分は何らかの形で集積され、メタデータの部分は別個付随する形で表記される。
- メタデータは、本体から演繹される部分と、本体とは別個の形で表される部分がある。

# 月震データベースの例

```
@@ MQ100 dbconv-1.00
Data_create_date:Thu Aug 25 19:10:26 1994 JST
Average_amplitude:        521.230498,522.601794,513.186227
Channels:         3
Data_format:      NEW
Maximum_amplitude:        534.000000,530.000000,515.000000
Data_type:        LP
File_type:        FULLTEXT
Number_of_data:   51840
Observation_mode:FLAT
Original_data_file:       /moa/se-1-1/file022
Sampling_rate:    0.150940
Station: AP12
Start_record:     83
End_record:       90
Start_time:       1972 155 6 10 46 496
Tape_number:      1001
@@
522.000000,520.000000,514.000000
522.000000,519.000000,514.000000
523.000000,519.000000,514.000000
523.000000,518.000000,513.000000
```

ヘッダ

データ

寺薗(1993)、Terazono(2007)

# 惑星科学データの基本

- 多くの惑星科学データは、いまのように、「ヘッダ＋データ」の形で表される。
  - 惑星画像をはじめとする惑星探査データの格納形式のPDS、天文データのFITS、流体力学データのNetCDFなど、ほとんどのデータそうといってよい。
- ヘッダとして表現されたデータは、本体のデータの要約と、**本体のデータからは再現されない**データの2種類に分かれる。
  - 本体のデータの要約は、迅速なデータ解析や、データそのものの要素の表現のために便宜的につけられるケースが多い(例えばデータの平均値など)。
  - 問題は、本体から再現されないデータ。これが、本来の意味のメタデータ。


# メタデータのデータベース化

**前述のような「ヘッダ＋データ」は、データベース化は比較的容易。**

- ヘッダ部分はテーブルとしての表現が可能。データ部分はラージオブジェクト、ないしはテーブルとしての表現が可能。
- 現在であれば、ヘッダ部分はXMLで記述ないしはやりとりし、半自動的に機械的解釈を行うことも可能である。

# メタデータは どこまでの範囲を指すか

**そもそも、メタデータというものの範囲について明確な定義、あるいはコンセンサスというものがあるのだろうか？**

- データを表現するためには、そもそもデータの構造そのものを表現する必要がある。
  - ヘッダ＋データ型のフォーマットは自明かもしれないが、それは将来にわたって担保されるものではない。
  - ヘッダの意味が何だかわからないというケースもあり得る。また、その数値が何を意味しているのかわからなければ解析にも何も役立たない。
- 広義のメタデータとして、こうした「データを取り巻く環境を記述するデータ」という概念が成り立つ。

# データは何が保証されるべきか

- アクセシビリティ
  - 大容量データの解析時代、科学者はなるべく簡便な手段でデータアクセスを行うことにより、手間を最低限に減らしていくことが必要。
- 永続性
  - いつのデータであっても、そのデータが得られた状況などをしっかりと把握できることが必要。
- データの説明
  - そのデータが何を意味しているのかが、明確な形で理解されなければならない。

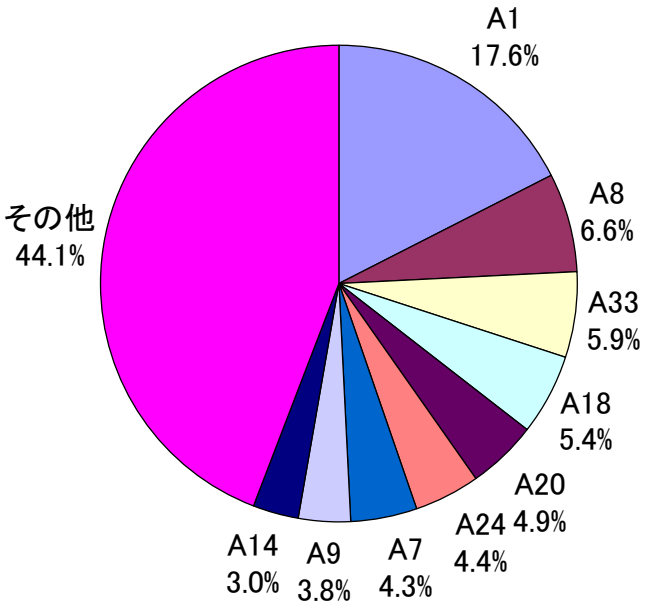

# データ記述性

- 1つの考え方として、XMLを応用して、データそのものの考え方を記述するという方式が考えられる。
- 形式としてはDTDを拡張する。ただし、DTDは1つのファイルに対して1つという形であるから、例えばヘッダごとにDTDをつけるということは難しい。
- DTDそのもの永続性が保証されるか？　名前空間が消滅してしまうとDTDも持たなくなってしまう。

# データマイニング的な考え方

- データマイニングの考え方でいうと、「メタデータがついていないデータから、様々な手法でメタデータ的な要素を取り出し、定義付けをしていく」ことが重要である。
- ただし、データマイニングの場合、その効率の問題もさることながら、統計的手法によって演繹された結果が本来のデータに対してどのような意味を持つのかという問題が含まれる。
- 地球惑星科学の場合、すでにたいていのデータはタグ付けされているため、データマイニングの手法はあまり適切ではない。むしろこの場合は「データエクスプロレーション」(data exploration)として、メタデータをさらにデータマイニング的に解析するという手法がとれる。

# 例えば(1)月震DB




月震データベースの中から、ヘッダ要素だけを使った解析例(メタデータ解析ともいえる)。ヘッダから、深発月震のグループを割り出し、それぞれのグループのデータ量を計算したもの。
これまで、このような統計例は(不思議なことに)存在しなかったが、これでみると、今までそれほど多いと考えられていなかったA18グループが意外に多いということが浮かび上がってくる。

非常に初歩的なデータマイニングだが、これは簡単なスクリプトを書くだけで可能になった事例。

Terazono (2007)

# 例えば(2)PDS画像

月探査画像の画質判定の手法として、

画像の最頻輝度値の比率(p)と、最頻輝度値と第2最頻輝度値の比率(r)が有効に利用できる。

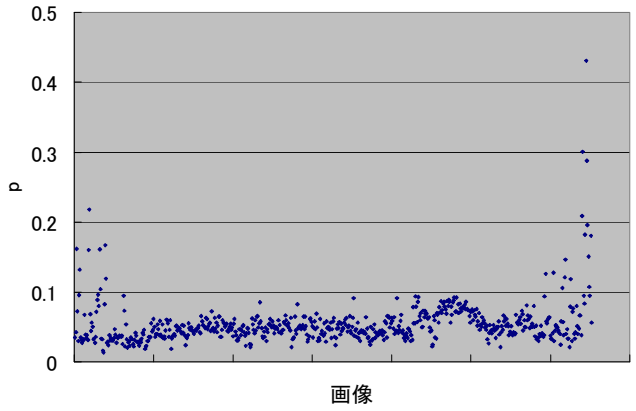
特に、pが小さくr が大きな画像は、コントラストもよく解析に適する画像が多い。

たとえば、右図から高緯度地域には画質がよい画像が多いことが、この図からも示唆される。

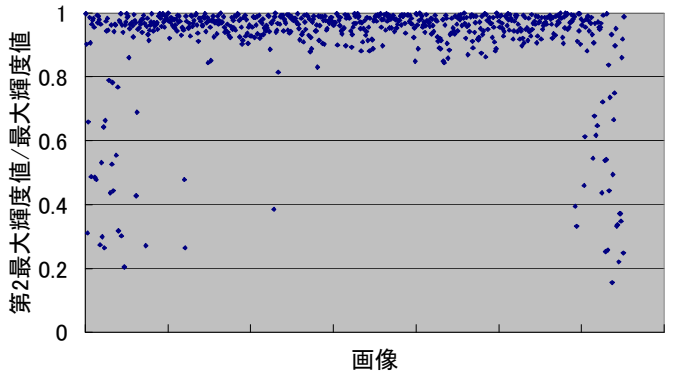
これらの成果は、実はスクリプトを書いてデータからパラメータを抽出してできたもので、ある種のメタデータ解析といえるであろう。

寺薗・齋藤 (1998D)

最頻輝度値の比率 (p)




最頻輝度値/第2最頻輝度値の比 (q)




南 → 北

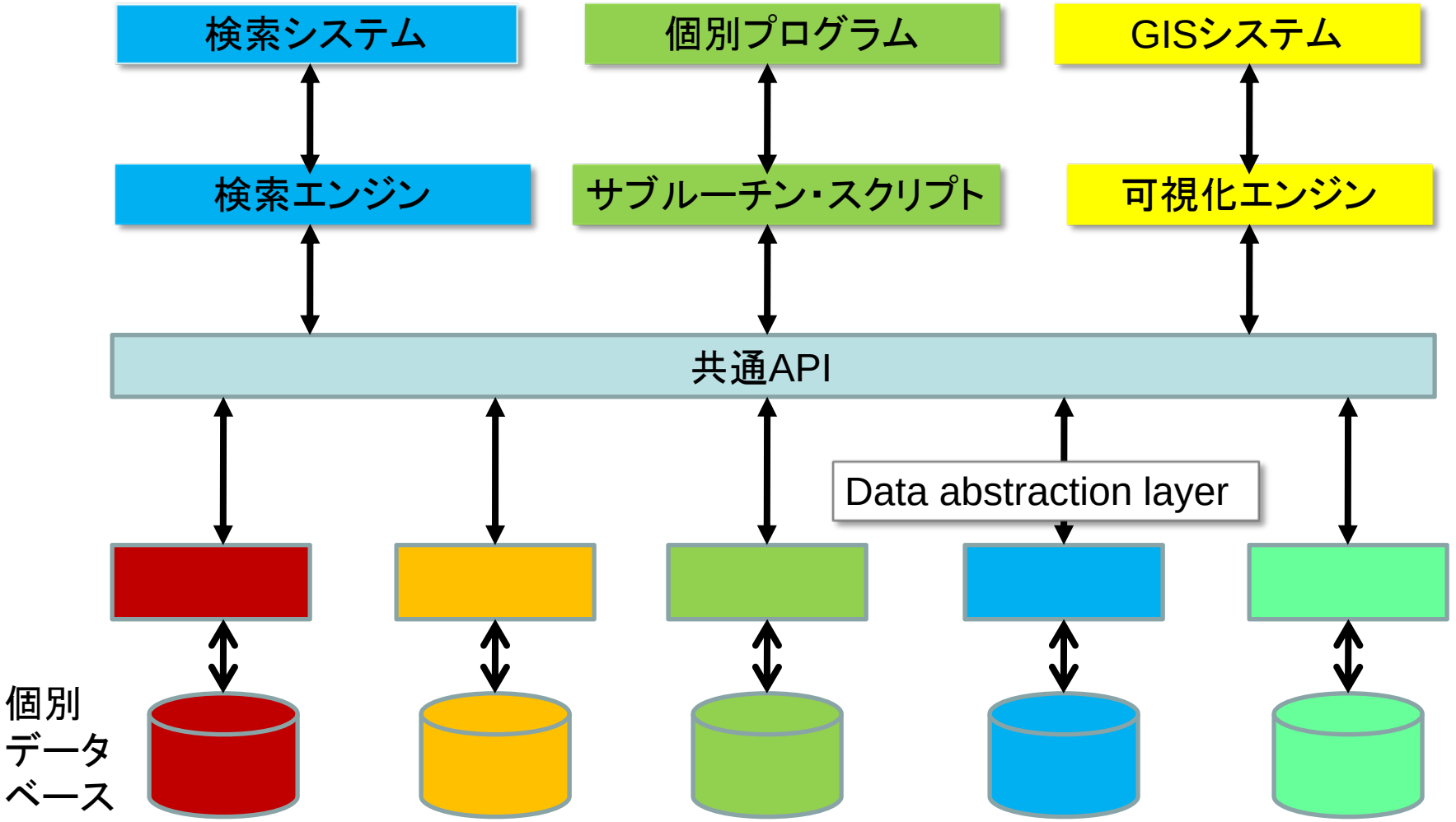

# データベースの考え方

- 現代のデータベースエンジンがアクセスのために用いている言語がSQLであり、その習得が難しい以上、科学者に対する何らかの統一されたフロントエンドが必要。
- たいていの場合、データベースは一度だけ構築すればよく、最終的にはデータウェアハウスという考え方に落ち着くことになる。
  - 従って、データの取り出し方と、そのデータに対する意味づけが重要になる。
- そのデータを記述するための統一された枠組みは存在するのか？　それがいまやいちばんの問題。
- さらに、存在したとして、それがデータベースとの親和性がよいかどうかにも気をつけた方がよい。

# ネットワークデータベース

- ネットワークで結ばれたデータベースを、科学者が意識せずに利用できるようなプラットホーム(地球惑星科学データウェアハウス)を構築していくことが必要。
  - 分野ごと、領域ごと、データの種類ごとに分かれている個々のデータベースに統一したI/Fでアクセスできるような枠組み(＝サブルーチン、フロントエンド、API、規約、…)
  - はやりはウェブベースかもしれないが、解析などにはやはり「ファイル」としてもらえることが必要。
  - 「小さく産んで大きく育てる」ポリシーで、まずは2研究機関、あるいは少数グループではじめていき、その後規格化を目指していく。

# 地球惑星科学データフレームワーク




寺薗・齋藤(2000)に基づく

# まとめ？

- まず、メタデータというのがどの範囲のものを指すかということを検討する必要がある。
  - 我々として押さえておくべき「メタデータ」の範囲はどこなのか。データに関係するところなのか、もっと広範囲なものなのか。
- メタデータのデータベース化は、地球惑星科学データ構造と深い関係を持つ。
  - 基本的な地球惑星科学データの構造はほぼ同じではあるが、問題はそれが、科学者の効率のよい解析につなげられるかどうか。
- データベース化と同時に、汎用的なアクセスツールの整備を検討する必要がある。
  - 地球惑星科学のデータウェアハウスのようなものを一元的に整備できれば、おもしろいことになるかも。