# 地球流体物理科学者のためのデータアーカイブサーバ構築支援ツール: Gfdnavi におけるデータベース設計と検索インタフェースの実装

# Metadata Schema Design and Query Interface for Gfdnavi: A Data Archiving Server Construction Support Tool for Geophysical Fluid Database

柳平 有美♥ 渡辺 知恵美♥
堀之内 武♠

Yumi YANAGITAIRA Chiemi WATANABE
Takeshi HORINOUCHI

**近年，地球流体物理科学データは爆発的に増加しており，科学者たちは自らが保有するデータから必要なデータを検索したり，科学者同士で互いに公開し合いたいという要求が高まっている．このような要求に応えるため，我々は科学者にかかるコストをできる限り削減することを目的とした地球流体物理科学者のためのデータアーカイブサーバ構築支援ツール：Gfdnavi の開発を進めている．本稿では科学データにおけるメタデータのスキーマ定義と，メタデータの自動生成，検索インタフェースの実装について述べ，さらには検索結果のグループ化とランキング手法について提案する．**

We develop Gfdnavi: a data archiving server construction support tool for geophysical fluid database. This system can cut the cost for scientists to construct data archiving server which services high functionalities for metadata search, analysis and visualization. In this paper, we describe about a metadata schema definition for scientific datasets, a automatic metadata extraction module, and a the query interface by using Google map. Particularly, in the query interface, we introduce some methods of grouping and ranking result data, the interface design can lead users to narrow search conditions to find their demanding data interactively.

## 1. はじめに

近年，地球観測の測定機器の高機能化と計算機の高性能化により，地球流体物理科学データは爆発的に増加しており，科学者たちは自らが保有するデータから必要なデータを検索したり，科学者同士で互いに公開し合いたいという要求が高まってきている．しかし，一般の科学者が自身でデータを検索・公開し合うためには作業コストや学習コストがかかる．そこで，低コストで尚且つ簡単に検索・公開を実現できるツールが必要とされている．このような要求に応えるため，我々は地球流体分野のライブラリ開発チームである地球流体電脳倶楽部[1] と共同で，地球流体物理科学者のためのデータアーカイブサーバ構築支援ツール：Gfdnavi の開発を進めている．本稿ではGfdnavi におけるメタデータ自動生成および検索インタフェースについて述べる．

♥ 学生会員 お茶の水女子大学理学部情報科学科
yumi@db.is.ocha.ac.jp
♥ 正会員 お茶の水女子大学大学院人間文化研究科
chiemi@is.ocha.ac.jp
♠ 京都大学生存圏研究所 horinout@rish.kyoto-u.ac.jp

## 2. Gfdnavi

我々が開発している地球流体科学者のためのデータアーカイブサーバ構築支援ツール:Gfdnavi は，地球流体科学者が個人で持つ膨大な科学データをローカルで検索したり，共同研究者や同分野の科学者同士でデータを公開し合いたいという要求に対し，それを簡単に実現することを目的としたツールである．Gfdnavi は Ruby on Rails [5] という Ruby Web アプリケーション開発フレームワークを拡張し，地球流体科学者を対象とした高機能かつ導入が容易なデータカイブサーバ構築を支援する．この Gfdnavi の特徴は以下のとおりである．

**(1) メタデータ自動抽出：**地球流体学者の標準的ファイルフォーマットを対象に自動的にメタデータを抽出する．

**(2) 高機能な検索・分析・可視化インタフェース**
検索から様々な分析・可視化までをサポートする高機能な公開サーバが実現できる

**(3) デスクトップアプリケーションとしての個人利用**
ローカルホストでサーバをあげることにより，デスクトップサーチとして個人利用することもできる．デスクトップ上のデータ管理から可視化，分析まで一連の処理を通して扱うシステムは有用である．

**(4) 他の Gfdnavi との連携**
Web サーバとして公開する他，他の Gfdnavi サーバとの横断的検索等の機能をもち，それぞれのサーバ間を横断的に検索してデータを共有することが出来る．

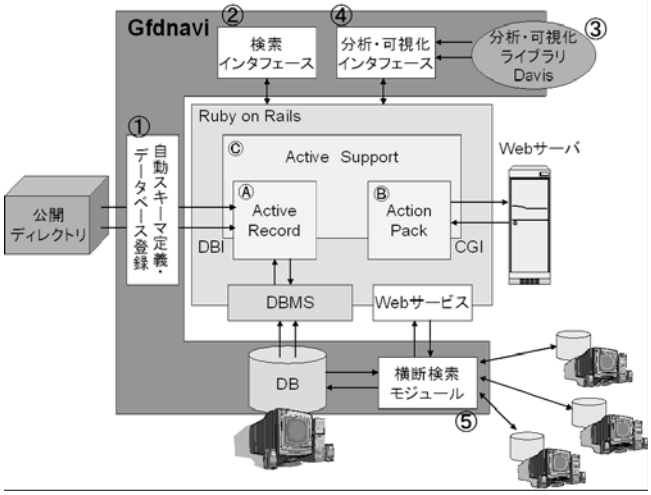


図 1 Gfdnavi の構成

Fig.1 System Configuration of Gfdnavi

Gfdnavi は以下の３構成に大別できる．

**(1) データ検索部**
公開用ディレクトリをスキャンしてデータファイルを自

動的に認識し，メタデータを抽出して登録する(図 1 ①)．さらに，GoogleMap を用いた高機能な検索インタフェースを提供し，ユーザが手間なく検索を行えるようにする(図 1 ②)．

**(2)データ分析・可視化部**

共同研究者の堀ノ内博士らは，これまで観測データをもとに分析・可視化するための Fortran ライブラリ DCL，Ruby ライブラリ Davis を提供している[4](図 1③)．これを Gfdnavi に搭載することにより，豊富なデータの分析・可視化を行えるようにする(図 1④)．

**(3) データ公開部**

P2P を利用して個々に立ち上げている Gfdnavi サーバを横断的に検索し，個々が持つデータを容易に共有することを可能にする(図 1⑤)．

本稿ではこのうち(1) のデータ検索部について述べる．Gfdnavi 全体の概要および(2) のデータ分析・可視化部については[3] を，(3) のデータ公開部については[4] を参照していただきたい．

## 3. データ検索部の開発

データ検索部は以下の構成よりなる．

**(1) スキーマの定義**

地球流体分野において広く使われるデータセットの構造に基づき，この分野において汎用的なスキーマを定義する．

**(2) メタデータの自動抽出**

ユーザの指定した公開ディレクトリ下にある科学データファイルをスキャンし，ファイルヘッダ等の情報から自動的にメタデータを抽出してデータベースに登録するツールを開発する．

**(3) 検索インタフェース**

空間情報・時間情報・キーワードを検索パラメタとし GoogleMap を用いたインタフェースを作成する．

### 3.1 メタデータ定義

科学データには衛星データやゾンデデータ，レーダデータ，シミュレーションデータなどがあり，1 つのデータセットに関連するメタデータ情報としては計測範囲（緯度・経度）や計測条件，計算モデル，計算パラメタなど多様な属性が考えられる．またファイルフォーマットも用途によって多種類存在するが，我々は NetCDF を参考に他のフォーマットでも柔軟に対処できるメタデータ定義を行った．NetCDF はメタデータを内包した自己記述型データフォーマットであり，図 2 に示すように一つのファイルの中に複数のデータセットが構成されている．

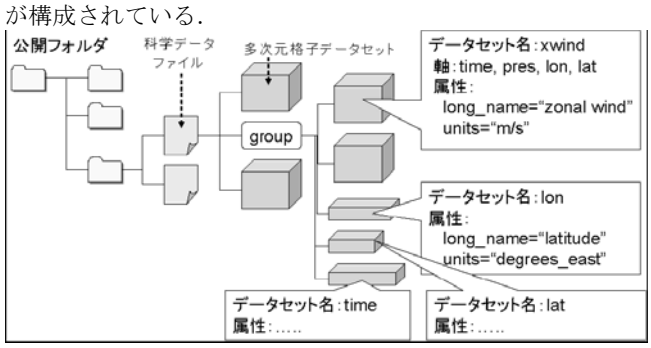


図 2 一般的な地球科学データファイルの構造

Fig.2 Data Structure of Typical Earth-Fluid Data File

この例では，ファイルの中に xwind という風速データ，気温データ，緯度データ，経度データなどが含まれている．各々のデータセットには複数の属性を付与することができる．属性にはデータ観測における条件やシミュレーションにおけるパラメタセットなどが記述されている．データセットは多次元配列であり，n 個の 1 次元配列を互いに直行する軸として定義することにより n 次元格子を定義することが出来る．図 2 中の xwind データセットは long name = ” zonal wind” units =”m/s” を属性として持つ．図 3 は NetCDF ファイル形式のヘッダ情報である．このヘッダ情報には”dimensions:”以下に格子の軸となる配列の配列名が示されており，variables:以下に多次元格子データの情報が示されており，このファイルには xwind, lon, lat, pres, time という多次元格子データ(variable) がファイル内にあることが分かる。また

*floatxwind*(*time, pres, lat, lon*)

という記述により多次元格子データである xwind が time（時刻）,pres（気圧），lat（緯度），lon(経度）という一次元配列データを軸に持つ float 型の 4 次元格子をしていることが読み取れる．また，

xwind：long name = ”zonalwind”

xwind：units = ”m/s”

は xwind に付与されている属性である．

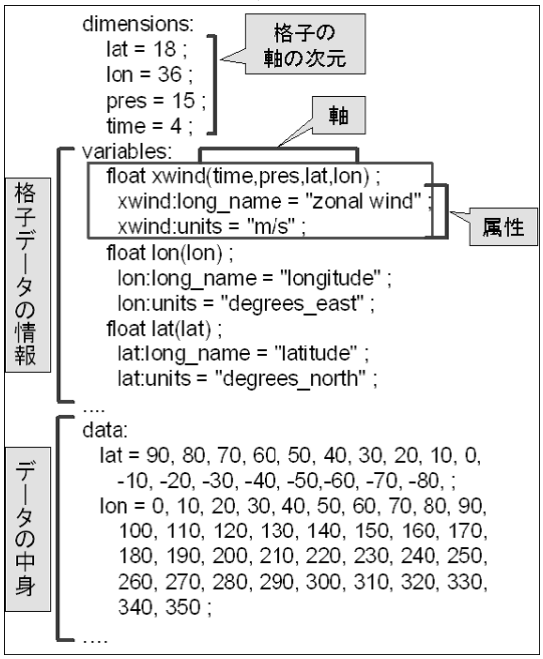


図 3 NetCDF のヘッダ情報例

Fig.3 An Example Header of a NetCDF-formatted File.

このようにして 1 つのファイルの中に複数の多次元格子を定義し，それぞれの格子データに属性を付与することが出来る．また多次元格子データの格子構造は現在サポートしている NetCDF に限らず，基本的に以下の 3 種類に分けることができる．

- **Grid 型**：緯度・経度方向に平行/垂直な軸を持つ．
- **Swath 型**：衛星スキャン方向に平行/垂直な軸を持つ．
- **Points 型**：任意の点集合で表される．

この分類および定義は NetCDF における気象データのためのフォーマット規約である CF 規約においてもほぼ同様の

格子に関する記述があり，またそのほかのデータ形式においても多少の違いこそあれ上記の格子構造のいずれかとして考えることができる．これらのことを元に以下のテーブルを定義した．

```
directories (id int,  parent id int,  name varchar,
             path varchar,  plain file tinyint)
variables(id int,name varchar,directory id int,
          path varchar)
spatial attributes(id int,variable id int,
          directory id,int,longitude lb float,
          latitude_rb float,longitude_rt float,
          latitude_rt float)
time attributes(id int,variable id int,
          start time datetime,end time datetime)
keyword attributes(id int,variable id int,
          directory id int,name varchar,value text)
```

本システムではデータセットのテーブルを variables とし，全ての科学データが持つ重要な属性として，空間属性(spatial attributes) と時間属性(time attributes) を抽出し，それら以外の属性値をキーワード属性(keyword attributes) として扱うこととする．また，それぞれのデータセットはフォルダごとにグループ分けされる場合もあるためディレクトリのテーブル(directories) を用意した．

**【空間属性】**

前述のとおり，格子構造は Grid, Swath, Points の 3 種類のタイプに分けることが出来る．これらの格子構造に対し，それぞれの空間属性は図 4 のように最小矩形の緯度･経度の最大値(lat rt,lon rt) と最小値(lat lb,lon lb) をとるように統一した．これにより格子構造に依存せずにデータセットの空間属性を表すことが出来る．

- Grid 型の場合（図 4(1)）：そのまま緯度・経度軸の最大値および最小値をとる．
- Swath 型の場合（図 4(2)）：長い帯状の格子を短い区間に区切り，各区間における緯度・経度の最大値と最小値をとる．この場合，1 つのデータセット(variable) に対して複数の空間属性が定義されることになる．
- Points 型の場合（図 4(3)）：緯度経度の最大値および最小値は同じ値となる．この場合も 1 つのデータセットに対して複数の空間属性が定義される．

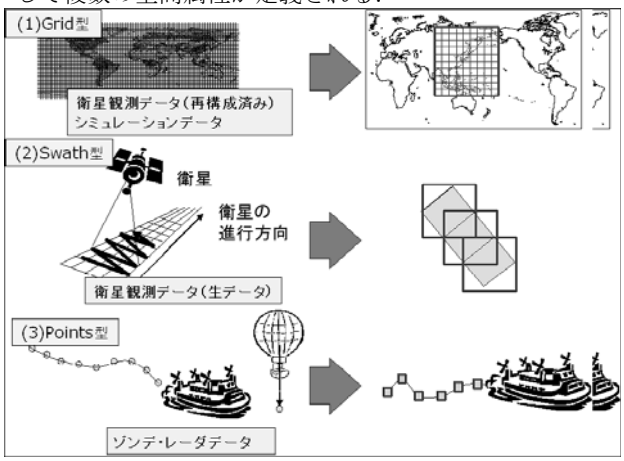


図 4 データセットからの空間属性の抽出

Fig.4 Spatial Attributes Extraction from Dataset

**【時間属性】**

時間属性は 0 次元または 1 次元の配列であるため，その最小値を start time，最大値を end time とする．

**【キーワード属性】**

キーワード属性はデータセットそのものの説明から，測定パラメタ，シミュレーションパラメタなど実に多種類の属性が存在するが，属性の定義はファイルに保存されているデータの種類やファイル生成者の方針によっても異なる．そこでメタデータとして保存する場合には属性値を特定せず，属性名と属性値のペアという最も単純な形で保存させることとした．

以上の方針に基づき，図 2 の xwind からは以下のメタデータが抽出されデータベースに定義される．

```
variable
    =(v001,xwind,NULL,"/filepath/filename:xwind/")
spatial attributes
    =(s192,v001,NULL,min(longitude),
      min(latitude),max(longitude),max(latitude))
time attributes
    =(t123,v001,min(time),max(time)
keyword attributes
    =(k023,v001,NULL,"long name","zonal wind"),
     (k024,v001,NULL,"units","m/s")
```

### 3.2 メタデータの自動生成

本モジュールは公開ディレクトリにある対象フォーマットのファイルをで読み込み，そこから前節に述べた方針に基づいてメタデータを抽出することによってメタデータを生成する．ここで，データセットと空間軸の関係から以下のように格子構造型を求める．

**Grid 型**：データセットの格子軸の中に緯度軸と経度軸が含まれている場合は Grid 型である．

**Swath 型**：Swath 型は衛星の進行方向に垂直な軸と平行な軸で格子が組まれるため，緯度経度は格子軸にはならない．この場合，緯度・経度データは衛星の進行方向に垂直な軸と平行な軸からなる 2 次元配列データとなり，緯度･経度データと同じ軸を含む多次元配列データセットの格子構造型は Swath 型となる．

**Points 型**：1 次元配列の緯度・経度データの配列数が同じで，かつそれらと同じ配列数を持つデータセットは Points 型である．

## 4. 検索インタフェース

検索インタフェースは図 5 のように，空間領域，時間領域，キーワードのどれかを指定すると画面の下部に検索結果のリストが表示される．検索結果が 100 件よりも多い場合は以下の手法で空間属性および時間属性を用いてグループ表示し，絞込み検索を促すインタラクションを設計した．


図 5 検索インタフェース

Fig.5 Search Interface

**【地図上でのグループ化表示】**
問合せに対する該当データを全てGoogleMap 上で表示すると繁雑になるため，それらをグループ化して表示させる．そして図6のように，GoogleMap をズームするごとにさらに細かいグループに分割して表示するようにする．
さらに，該当データのリストとGoogleMap の表示をリンクさせ，GoogleMap 上で欲しいデータをクリックするとリストの該当するデータにチェックが入るようにし，逆にリストの方でクリックするとMap 上の該当する部分の色を変えるようにする．

**【緯度円・経度円データの扱い】**
科学データの中には同一緯度(経度) 円上のデータの平均をとったデータもある．このデータには緯度情報もしくは経度情報しか含まれていない．そこで，そのようなデータに対しては，図7 のように点データと同じようなアルゴリズムでグループ化を行い，そのグループに含まれるデータの件数をGoogleMap の端に表示する．さらに，あるグループをクリックするとそのグループを細分化してデータのリストを表示するようにする．

**【時間属性によるグループ化】**
時間属性に関しては全球データも検索の対象とし，空間属性に対する検索結果と同様にグループ化も行う．検索結果は図8 のように時間軸を一定区間ごとに区切り，それぞれの区間に該当するデータ数を表示する．さらに，時間属性で絞込んだデータの中から空間属性についての検索ができるように，時間属性においても点データと部分データ，全球データは分類して考える．

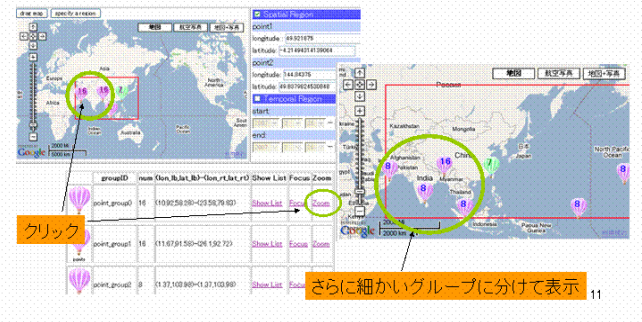


図6　空間属性による結果のグループ表示

Fig.6 Grouping Query Results by Spatial Attributes

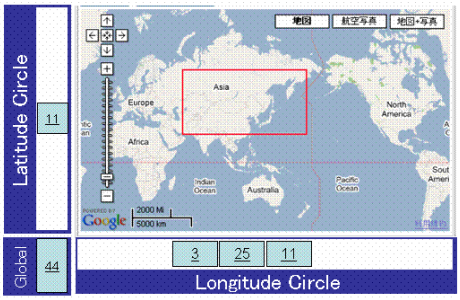


図7 緯度円，経度円データの表示

Fig.7　Showing equi-Longitude/Latitude Circle Data

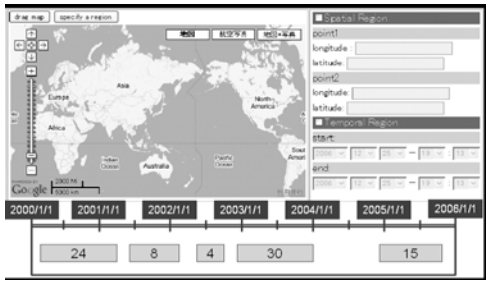


図8 時間属性による結果のグループ表示

Fig.8 Grouping Query Results by Time Attributes

## 5. まとめと今後の課題

本稿では地球流体物理科学者のためのデータアーカイブサーバ構築支援ツール：Gfdnavi のシステム構成のうち，データ検索部における汎用的なメタデータのスキーマ定義，メタデータの自動生成法，そして検索を促す検索インタラクションについて述べた．今後は提案した検索インタフェースのインタラクションを実装し，提案手法の妥当性についての考察を行う予定である．

## [謝辞]

本研究は，文部科学省科研費特定領域「情報爆発時代に向けた新しい IT 基盤技術の研究」の課題 A01-14（課題番号18049043）により行われた.本研究遂行にあたって様々な協力やコメントを頂いた西澤誠也，森川晴大，林祥介，塩谷雅人氏ら地球流体電脳倶楽部の各氏に感謝する．

## [文献]

[1] 地球電脳倶楽部 http://www.gfd-dennou.org/
[2] Chiemi Watanabe: “地球惑星科学研究者のためのデスクトップサーチツールの開発に向けて,” 情報処理学会研究報告 2006-DBS-140(2), Vol.2006, No.78, pp.429-436, 2006.
[3] 堀之内武，西澤誠也，渡辺知恵美，森川晴大，神代剛，林祥介，塩谷雅人“地球流体データベース・解析・可視化のための新しいサーバー兼デスクトップツール Gfdnavi の開発”，データ工学ワークショップ(DEWS2007) , D2-8 (2007)
[4] 佐藤麻美，渡辺知恵美“P2P を利用した地球流体データの横断検索・共有システムの実現に向けて”，データ工学ワークショップ(DEWS2007) , D1-9 (2007)
[5] Ruby on Rails http://www.rubyonrails.com/

**柳平 有美　Yumi YANAGITAIRA**
2007 年お茶の水女子大学理学部情報科学科卒業．科学データベースシステムの研究・開発に従事．日本データベース学会学生会員．

**渡辺 知恵美　Chiemi WATANABE**
お茶の水女子大学大学院人間文化創成科学研究科講師．博士(理学)．2003 年お茶の水女子大学大学院人間文化研究科博士後期課程修了．高度データベース応用の研究・開発に従事．情報処理学会会員．日本データベース学会会員．

**堀之内 武　Takeshi HORINOUCHI**
京都大学生存圏研究所助教．博士(理学)．1997 年京都大学大学院理学研究科博士後期課程修了．気象学の研究および気象学のための開発に従事．情報処理学会・日本気象学会・米国地球物理学連合・地球電磁気地球惑星圏学会会員．