# Gfdnavi:地球流体データアーカイブサーバ構築支援ツールの開発

柳平 有美　(指導教員：渡辺 知恵美)

## 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．一般に大きな企業や団体，例えば NASA や NOAA ではデータセンタを設置して数百 TB 若しくは数 PB の膨大なデータを管理し，Web 上でデータを公開している．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，我々は京都大学の堀之内武博士を中心とした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．

Gfdnavi は大きくデータ検索部とデータ分析・可視化部，データ公開部の 3 つに分けられ，本稿では Gfdnavi の構成とデータ検索部について述べる．

## 2　Gfdnavi

Gfdnavi は Ruby on Rails[4] を拡張し，地球流体科学者を対象とした高機能なデータカイブサーバ構築を支援するパッケージである．

Ruby on Rails とは，Ruby 言語による Web アプリケーション開発フレームワークで，データベースとのやり取りを行う ActiveRecord(図 1Ⓐ) と Web サーバとのやり取りを行う ActivePack(図 1Ⓑ)，そしてそれらを連携させる ActiveSupport(図 1Ⓒ) で構成される．

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ができる．

Gfdnavi の構成は以下の 3 つに大別できる．

**(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 1①)．さらに，GoogleMap を用いた高機能な検索インタフェースを提供する (図 1②)．

**(2) データ分析・可視化部**

共同研究者の堀ノ内博士らがこれまで開発してきた，観測データをもとに分析・可視化するためのライブラリ [2](図 1③) 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)．

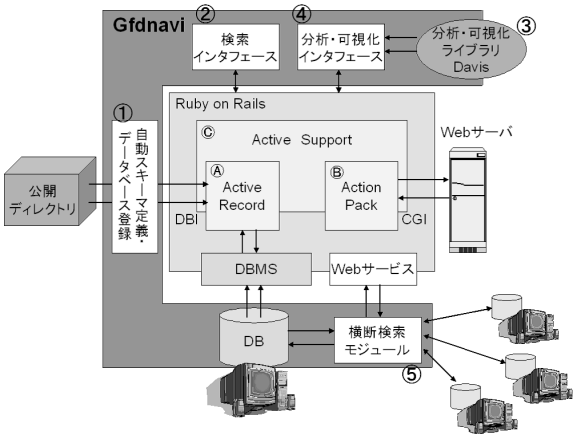


図 1: Gfdnavi の構成

次節より，このうちのデータ検索部について詳しく述べる．

## 3　データ検索部の開発

データ検索部では，科学データのメタデータベース化および検索インタフェースの開発にかかるコストを出来る限り削減することを目的とし，サーバ上の科学ファイルのメタデータの自動抽出モジュールと，機能豊富な検索インタフェースを提供する．これによりユーザは，Gfdnavi を起動するだけでサーバ上の科学データを検索できるようになる．また，Gfdnavi をローカルサーバとすればデスクトップサーチとしても利用出来る．

### 3.1　メタデータの定義と自動抽出モジュール

本モジュールは地球流体科学分野において現在主流となっている NetCDF というファイル形式を対象とし，メタデータ抽出を行う．

NetCDF と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，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．また，地球流体データセットはその特徴として，軸または属性値の中に空間情報と時間情報を含んでいる．

これらのことを元に図 2 のような E-R ダイアグラムを設計し，テーブルを定義した．

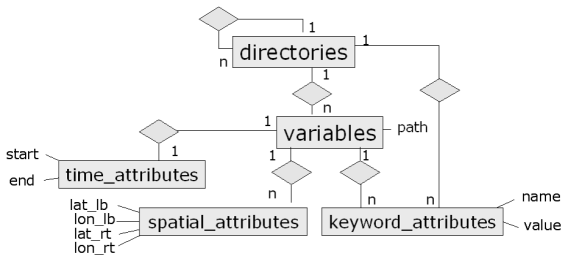


図 2: E-R ダイアグラム

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 (keyword_attributes) として扱うこととする．

メタデータの自動生成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読み込み，そこからメタデータを抽出して生成する．

## 4 検索インタフェース

検索インタフェースは図 3 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．また，空間領域に対しては GoogleMap を用いて検索したい領域をドラッグするだけでも指定できるようにした．

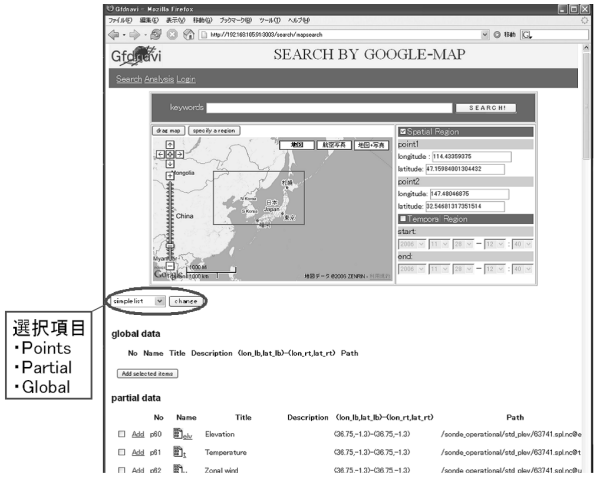


図 3: 検索インタフェース

さらに，より効率的に検索結果を表示させるため，検索結果が 20 件より多い場合はリストを表示せず，検索を繰り返してデータを 20 件以下に絞り込んだところでリスト表示させることとした．

### 4.1 空間属性に対する検索

検索においてはそれぞれのデータを，ある 1 地点の点データと，ある程度の範囲を持つ部分データ，そして地球全体をカバーする全球データに分類する．ここで，全球データは空間属性ではデータを絞込むことはできないため，時間属性とキーワード属性で検索をする．よって空間属性に対する検索においては，点データと部分データについてのみ考えることとする，

検索結果の表示については図 3 のように，GoogleMap の下に設けた選択切り替えを使用することで全球データ，部分データ，点データを分けて表示できるようにする．

また，問合せに対する該当データの数が多い場合，その 1 つ 1 つを全て GoogleMap 上で表示すると繁雑になるため，図 4 のようにそれらをグループ化して表示させ，ズームアップするごとにグループを細分化して表示するようにする．

### 4.2 GoogleMap 上のグループ化

問合せ該当データは，最初に指定領域の中心に近い順にソートし，その後グループ化を行う．

例えば点データのグループ化の場合，指定した問合せ領域の境界線のうち長い方の辺を e として，e/7 を閾値として扱うこととする．ある基準のデータからこの閾値内に含まれるデータは全て同じグループと見なし，含まれなければまた新たにグループを作成する．このようにして全てのデータにおいてグループ化を行う．

また，部分データに対してもほぼ同様にしてグループ化を行う．

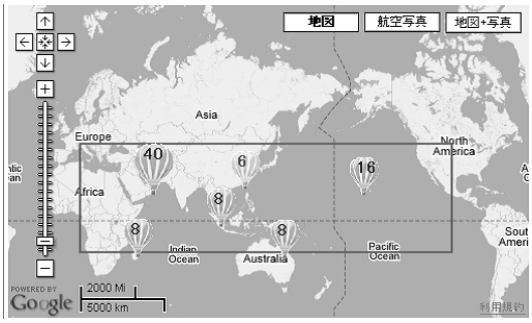


図 4: グループ化

### 4.3 ランキング

検索結果についてはランキングを行い，適合率の高いものから順に表示する．このランキングには以下の 2 つの計算を取り入れることとした．

(1) 中心点の距離問合せ領域の中心点とデータの中心点の距離を計算する．これは点データに対して特に有効になる．

(2) 問合せ領域に対するデータ領域の該当率問合せ領域に対してデータの領域がどれだけ重なっているかを計算する．これは部分データに対して有効になる．

## 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メタデータのスキーマ定義と自動生成法，時間・空間情報を重要視した検索インタフェースについて述べた．

さらに今後は，空間に関する検索インタフェースを充実させ，時間やキーワードに関するインタフェースも実装していきたい．

## 参考文献

[1] 地球電脳倶楽部http://www.gfd-dennou.org/

[2] 電脳 davis プロジェクト http://www.gfd-dennou.org/library/davis/index.html/

[3] 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.

[4] Ruby on Rails http://www.rubyonrails.com/