Gfdnavi: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 データ属性の探索的検索を利用する検索

重川美咲子† 西澤 誠也†† 堀之内 武††† 渡辺知恵美††††

† お茶の水女子大学理学部情報科学科 〒 112-8610 東京都文京区大塚 2-1-1
†† 京都大学数理解析研究所 〒 606-8502 京都府京都市左京区北白川追分町
††† 京都大学生存圏研究所 〒 611-0011 京都府宇治市五ヶ庄
†††† お茶の水女子大学大学院人間文化創成科学研究科 〒 112-8610 東京都文京区大塚 2-1-1
E-mail: †misako@db.is.ocha.ac.jp, ††seiya@kurim.kyoto-u.ac.jp, †††horinout@rish.kyoto-u.ac.jp, ††††chiemi@is.ocha.ac.jp

あらまし 我々は地球流体科学者を対象に，個々の科学者が保有するデータを低コストで尚且つ簡単に検索し，科学者間で共有するためのツール Gfdnavi の構築を行っている．Gfdnavi では科学データファイルに付与するメタデータの属性に対して属性名などの規定を行っていないため，公開者によって属性名が異なる場合があり，利用者は所望するデータを検索するためのキーワードに何を指定してよいかわからない場合がある．そこで我々は Gfdnavi の検索インタフェースに Faceted Navigation を採用し探索的検索を可能にした．本稿では現在開発中の Gfdnavi 用探索的検索インタフェースについて述べる．どのようにすれば容易に検索が行えるのかを考慮してデータを属性によりグルーピングし，検索インタフェースを設計した．また本検索機能を Web サービスにて公開することによって，利用者の解析プログラムからの利用を可能にする．

キーワード 科学ＤＢ，ユーザインタフェース

Gfdnavi: a package for supporting construction of scientific data archiving server Search using Exploratory search for attributes of data

Misako SHIGEKAWA†, Seiya NISHIZAWA††, Takeshi HORINOUCHI†††, and Chiemi WATANABE††††

† Department of Information Sciences, Faculty of Science, Ochanomizu University
†† Research Institute for Mathematical Science, Kyoto University
††† Research Institute for Sustainable Humanosphere, Kyoto University
†††† Graduate School of Humanities and Science, Ochanomizu University
E-mail: †misako@db.is.ocha.ac.jp, ††seiya@kurim.kyoto-u.ac.jp, †††horinout@rish.kyoto-u.ac.jp, ††††chiemi@is.ocha.ac.jp

Abstract We are developing Gfdnavi, which is a data archiving support package for geophysical fluid data. In this paper we describe a new query interface for Gfdnavi. Gfdnavi provides a simple database schema that every attributes of variables as a pair of attribute name and its value. Furthermore we don't have any rule for naming of attribute. Then users of Gfdnavi query interface may confound what keyword they specify for finding variables they want. To solve this problem, Gfdnavi query interface applies faceted navigation mechanism for querying data inductibely. Based on the metadata schema, we categorize attributes of metadata, and we design an appropriate query interaction. Additionally, we provide exploratory query functions as web service api to utilize them from the user program in a remote host.

Key words e-science, user interface

1. はじめに

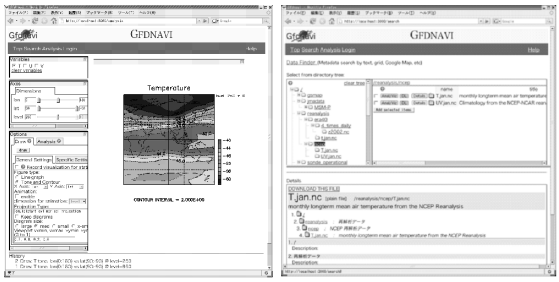
近年，地球観測の測定機器の高機能化と計算機の高性能化により，地球流体データは爆発的に増加しており，科学者たちは自らが保有するデータから必要なデータを検索したり，科学者同士で互いに公開し合いたいという要求が高まってきている．一般に大きな企業や団体，たとえば NASA [6] や NOAA [7] ではデータセンタを設置して数百 TB 若しくは数 PB の膨大なデータを管理し，Web 上でデータを公開している．しかし，一般の科学者が自身でデータを検索・公開し合うためには作業コストや学習コストがかかる．そこで，低コストで尚且つ簡単に検索・公開を実現できるツールが必要とされている．このような要求に応えるため，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 [3] 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．

様々な機能のある Gfdnavi の中でも我々はデータ検索部を担当している．本稿ではまず，データ属性の探索的検索を利用する新しいインタラクションモデルを採用したデータの検索方法について述べる．更にその中で使われる，検索インタフェースにおけるデータ属性のリストの表示について詳しく報告する．そして検索結果を取り扱うキャッシュ機能について述べ、最後に Web サービス化の提案をする．

2. Gfdnavi

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，地球流体物理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 上では科学データを検索・可視化・分析することが可能であり，我々はデータ検索部の開発を行っている．



可視化画面　　検索画面(tree view)

図 1　Gfdnavi の可視化画面と検索画面

また，Gfdnavi はローカルサーバで自分用に自分のデータを扱うことが可能な一方，Web サーバ上で Web サイトとして利用することも可能である．

2.1 データ構造

Gfdnavi で扱う科学データには図 2 のように以下の 2 種類の属性が定義されている．

（1） 空間属性と時間属性

（2） キーワード属性

空間属性と時間属性の二つは，どのデータにも共通で付与されている属性である．そして，キーワード属性とは空間属性・時間属性以外の属性をまとめたものである．Gfdnavi ではデータ属性のネーミングについて規定されていないので，属性名がデータ公開者次第になってしまう．そこで公開者によって異なる名前を定義をされている属性をキーワード属性とまとめているのである．

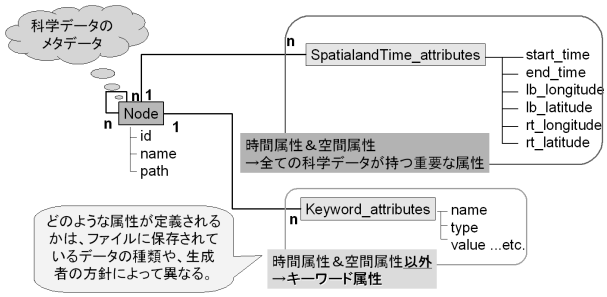


図 2　科学データのメタデータの属性定義

また，図 2 にあるように Node 同士は n 対 n の関係にある．つまり図 3 のように複数の Node がディレクトリ構造となっている．その様な Node のうち，自分の子として他の Node を持つものを Directory 便宜上と呼ぶことにする．また，一番下の階層の葉の部分にあたるものには Variable や image，function といったものがあたる．

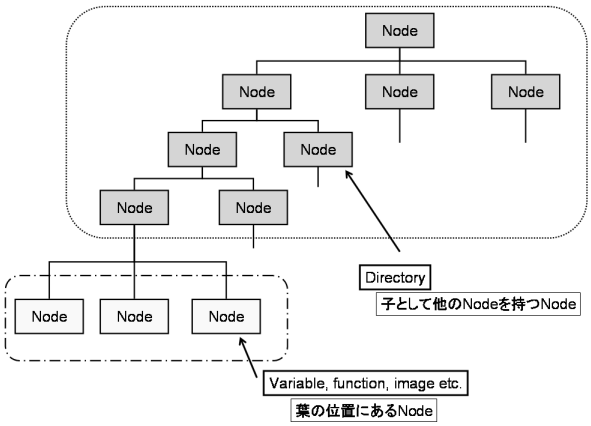


図 3　Node のディレクトリ構造

2.2 検索インタフェースの問題点

Gfdnavi では前述の様に，全ての科学データが持つ属性として時間属性・空間属性，そしてそれ以外の属性値をキーワード属性として，各科学データに 3 種類の属性を定義している．そして検索インタフェース（図 4）は，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ものとなっている．空間領域は緯度経度を数値入力で指定するほかに，GoogleMap での範囲の指定も可能である．

しかし，この先行研究 [1] のバージョンの検索インタフェー

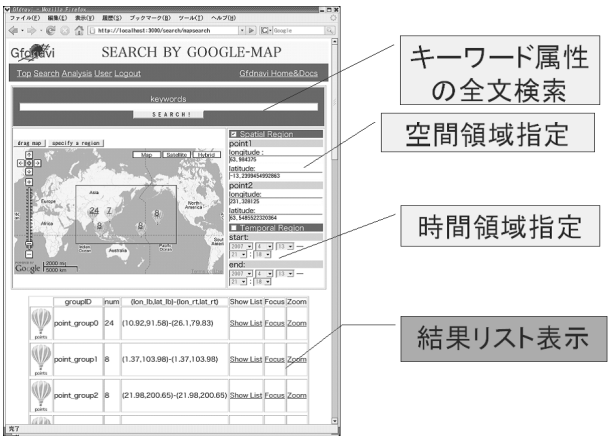


図 4　先行研究の検索インタフェース

スには改良すべき点が二点ある．

（1）　キーワードによる検索がユーザ入力形式

Gfdnavi では科学データファイルに付与するメタデータの属性に対して属性名の定め方の規定がない．従って公開者によって属性名が異なる場合がある．その様な場合，ユーザは必要なデータを探し出すために絞込み条件として使用すべきキーワード属性がわからない場合が考えられる．よってこのままではキーワード検索を有効活用しにくい．

（2）　空間領域以外の条件では絞込み検索を重ねることができない

必要とするデータを取得するためには，いくつかの絞込み条件を組み合わせた検索が必要な場合が多い．従って様々な条件での繰り返し絞り込み検索が不可能であると，ユーザが求めるデータにたどり着けない可能性が高い．

ユーザにとってのより簡単なデータの検索のためには，この二点を中心とした改良が必要であると我々は考えた．これらを踏まえ，新しい検索インタフェースの実装を進めている．以下，新しい検索インタフェースについて述べる．

3.　探索的検索を利用する新しい検索インタラクションモデルの実装

ユーザにとってのより簡単なデータの検索のためには，前述した二点を中心とした改良が必要である．

3.1　探索的検索

求めるものが明確にどのようなものかはわかっていない場合，データ全体に対して 1 回の検索をかけるだけでまさに必要としているものを得るのは難しいことが多い．従って大抵の場合我々は図 5 の様に絞込み検索を繰り返したり，検索結果を比較したり，更には検索結果同士を組み合わせたりしながら求めるものに近づいていく．この様に，試行錯誤しながら徐々に必要とするものに近づいていく検索方法を本稿では「探索的検索」と呼ぶ．

探索的検索の代表的な手法に Faceted Navigation がある．絞込み条件の選択肢としてあらかじめ設定されている属性値を，結果を見ながら変えるだけで容易に探索的検索を行うことが可能である．ユーザは自分で問合せ文を記述する必要がなく，システムにナビゲーションされる様にして簡単に検索が行える．検索結果を見て検索に使った絞込み条件の順番を変えること，

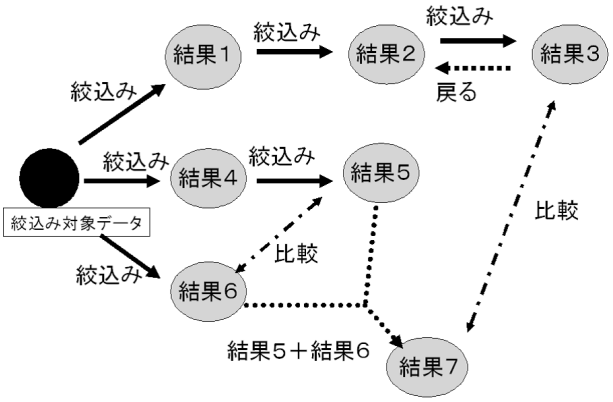


図 5　探索的検索のイメージ

一連の検索の中である属性で絞り込んだという条件のみ取り消したりすること等を可能にすることでユーザに探索的検索を促す．そしてこれを基にした検索システムが既にいくつか開発されている．

その良い例に Flamenco Search [4] がある．図 6 がそのインタフェースであり，ノーベル賞受賞者のデータに対して受賞した年や部門等で絞込み検索を行うことができる．インタフェースの左側に検索の選択肢となる属性値が表示されており，これらをクリックするとそれを絞込み条件とした検索が行われる．また使用した絞込み条件は検索結果の上部に表示されており，その隣にある×マークをクリックすることで検索からその条件のみを取り消すことも容易にできる．この様な Faceted Navigation に基づき，我々は Gfdnavi の新しい検索モデルの構想を立てた．

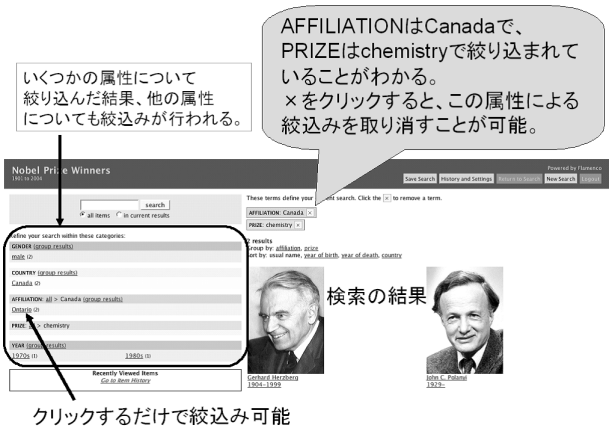


図 6　Flamenco Search

新しい検索インタフェース（これ以降 Gfdexplorer と呼ぶ）は図 7 に示す新しい検索インタラクションモデルを採用している．

top ページからの絞込みを最初の絞込みと見なし，次に top ページを開くまでに行った絞込みまでを一回の検索セットとする．一回の検索セットの中で絞込みに使われた条件を，一つのクエリセットとしてまとめて記録する．

検索の流れは以下のようになる．

（1）　データを属性によって分類・グルーピング化し，属性リストとしてユーザに対して表示する．

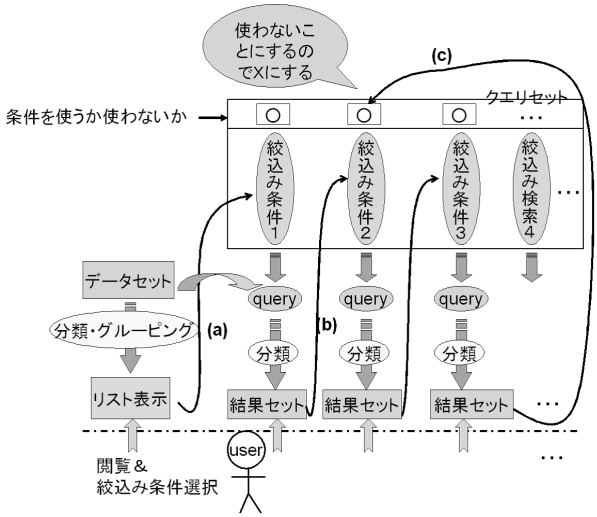


図 7　探索的検索を用いた新しい検索モデルのインタラクションイメージ

（2）　ユーザがそのリストから絞込み条件を選択すると，クエリセットにその条件が追加される（図 7 中 (a)）.

（3）　選択された絞込み条件によってデータセットに対してクエリを発行し，得られた結果を再び分類してリスト表示する.

（4）　得られた結果を見て，ユーザが更に絞込み条件を選択する．選択された条件はクエリセットに追加される（図 7 中 (b)）.

（5）　今まで選択した条件全てを満たすデータを結果として返し，また分類してユーザにリストを示す.

（6）　以上の手順を繰り返していくと複数の条件での絞り込みができる.

（7）　また，得られた結果を見てユーザが以前使ったある条件を取り消すということもある（図 7 中 (c)）．その場合はクエリセット中の取り消す条件以外の条件を満たす検索結果を返す.

（8）　その後も更にクエリセットに条件を追加したり，逆に取り消したりしながら検索を進めていく.

この新しいインタラクションモデルにより，ユーザはリスト中から属性値を選んでいくだけで簡単に必要とするデータを取得することができる．また，クエリセット中の絞込み条件から特定の条件を取り消すことができるので，容易に検索結果の比較を行うことが可能である.

以下，この検索モデルを実装した後の各属性部分と，ユーザに検索のヒントであるリストを与える為に新たに追加した機能について述べる．前提として，全ての属性について探索的検索が可能となっている．また，図 8 に Gfdexplorer を示す.

3.2　キーワード属性

キーワード検索において，既存のインタフェースのユーザ入力形式と併せてリストから選択する形式も採用した．絞り込み対象である各データのキーワード属性すべてをリストアップする．ユーザはリストアップされている値をクリックして選択すると絞込み検索が行われる．リスト部分も更新され，今絞り込んだキーワード以下，更にどのようなキーワードで絞り込めるかをまたリスト表示する．既に開いてある属性値リストが不要になった場合，[CLOSE] を選択することでそのリストを隠すことが可能である.

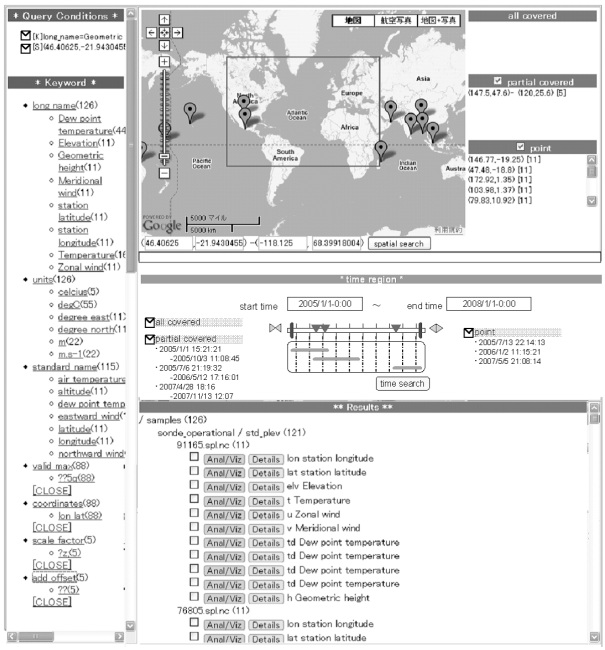


図 8　新しい検索インタフェース：Gfdexplorer

また，Top3 のキーワード属性については使用頻度が高いと考え，常にその属性の属性値をリスト表示することとした.

図 8 で Gfdexplorer を利用したキーワード検索の流れを紹介する.

（1）　gfdexplorer 左側のキーワードリストからキーワード属性を選択する.

（2）　選択したキーワード属性の真下に，その属性の持つ属性値一覧がリスト表示される.

（3）　その中から属性値を選択する.

（4）　もし開いた属性値リストが不要な場合は [CLOSE] で閉じて 1 の手順に戻る.

（5）　3 で選んだものを絞込み条件とした検索結果が gfdexplorer 下部の Results 部分に表示される．また，後述する Query Conditions の部分にどのような条件で絞込みを行ったかが表示される.

3.3　空間属性

空間属性による検索に関してはデータは領域を持つので，図 9 にもあるように，表示する際に次の 3 種類に分けた.

（1）　all covered：指定した領域全体をカバーする，領域を持つデータ

（2）　partial covered：指定した領域を部分的にカバーする，領域を持つデータ

（3）　point：指定した領域内に含まれる点データ

Gfdexplorer の GoogleMap 上では，指定した領域全体をカバーするデータは常に表示されるようになっている．そして”all covered”，”partial covered”，”point”を選ぶとそれぞれ部分的にカバーするデータ，点データが地図上に現れる．ユーザが GoogleMap 上で検索する範囲を指定するとその部分がズームアップされ，その範囲内にあるデータがまたどのようにユーザが指定した領域と適合しているか別に表示される.

キーワード検索のときと同様，空間属性における絞込み検索も Query Conditions に反映される．そこには GoogleMap 上

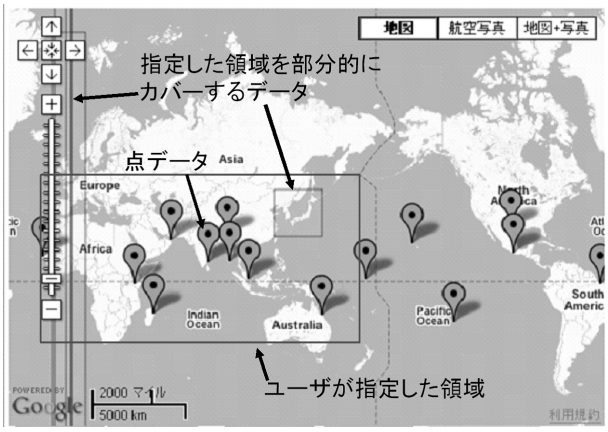


図 9　空間属性における検索

でユーザが指定した範囲の緯度経度が表示される．

3.4　時間属性

GoogleMap の下にあるバーで start time と end time を指定し，時間の範囲の絞り込みを行う．また空間属性と同様に領域を持つデータなので，まずデータを以下の 3 種類に分けた．

（1）　all coverd：選択した時間領域全体をカバーするデータ

（2）　partial covered：選択した時間領域を部分的にカバーするデータ

（3）　point：選択した時間領域内に存在する点データ

バーで絞り込んだ時間領域の中に属するデータを，点データはバーの上に，部分データはバーの下に検索結果として表示する．またバーの左右の＋と－のアイコンのクリックにより，絞込み対象領域を広げたり狭めたりすることも可能である．

前述した二種類の属性での検索と同じく，Query Conditions に使用した検索の条件が表示される．

3.5　検索条件：Query Conditions

検索条件の編集を行うことができる．まず絞込み検索を行うと，Gfdexplorer の左上にどのような条件で絞込みを行ったかが表示される．キーワード検索なら [K]，空間属性での検索なら [S]，そして時間属性における検索ならば [T] が頭に表示される．また，これは同じ回の検索中の絞込みを全て表示する．

各条件の左にあるボックスのチェックをはずすと，その条件を絞込みの条件からはずした結果を得ることができる．また，ボックスにチェックを入れると再び条件として追加できる．

よって一度複数条件で絞り込んだ結果からある条件を除いた検索結果や，ある検索結果からいくつかの条件を他の条件に入れ替えた際の検索結果を容易に導くことができる．

これらは絞り込んだ順番関係なく可能なので，容易に検索結果の比較を行うことが可能であり，ユーザがより簡単に求めるデータに近づくことができると考える．

3.6　検索結果の表示：Results

Gfdexplorer 下部に，絞込み検索を行った結果が表示される．そこで ”Anal/Vis ”のアイコンを選べば解析と可視化のページに，”Details ”のアイコンを選べば詳細表示に飛ぶことが可能である．これらのページについては論文 [2] を参照されたい．

また，ここではデータのパスを表示している．データは意味のある Directory にまとめられている場合がほとんどであるので，Directory 別にパス表記を行っている．

図 10(a) に，ある検索を行った際の Results 部分を示す．検索結果の上部に実線の円で囲んであるようにパスが表示されている．これから全データの中の”samples”Directory に検索結果となるデータが存在するということがわかる．更にその下の点線部分を見ると，”samples”Directory の下の適合するデータが存在する Directory が表記されている．図 10 の (b) では適合するデータの存在する Directory に色がついている．(a) の図での点線部分一つ目（＊1）は (b) の図での左側の 3 段目の階層の”std_plev”Directory に，(a) の点線部二つ目（＊2）は (b) の右側の上から 4 段目の階層の”2006”Directory にそれぞれあたる．

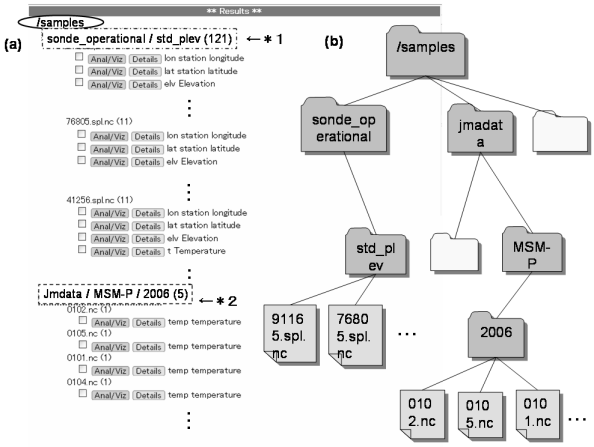


図 10　Results のパス表記

また，(a) と (b) の二つの図を比較してわかる様に，分岐したあと同じ階層に自分以外適合するデータを持つ Directory が存在しない場合，パスはまとめて”/”で区切ることで表記される．

3.7　クエリヒストリ

データベースの中に，query_histories というテーブルを用意した．その中身を図 11 に示す．但し図では見易さの為に表として表示する．構成は以下のようになっている．

id	querytype	user_id	time	description	queryset_id	conditions	count	cashed
1	keyword	1	2007-12-30 18:16:46	[K]long_name= Dew point temperature	1	name = 'long_name' and value= 'Dewpoint temperature'	15	1
2	space	2	2007-12-31 11:17:32	[S](88.59375, -33.72433966174759) -(258.75, 47.517200697839414)	2	longitude_lb <= 258.75 and longitude_rt >= 88.59375 and latitude_lb <= 47.517200697839414 and latitude_rt >= -33.72433966174759 and not(node_id IS NULL)	63	0
3	keyword	1	2007-12-31 16:27:52	[K]long_name= station latitude	3	name = 'long name' and value = 'station latitude'	15	1
4	keyword	1	2007-12-31 16:28:47	[K]standard_name= latitude	3	name = 'standard_name' and value = 'latitude'	15	1
5	keyword	1	2007-12-31 16:29:06	[K]units=degree north	3	name = 'units' and value = 'degree north'	15	1

図 11　query_histories の中身

- id：検索の id
- query_type：検索の種類（キーワード，時間，空間のいずれか）
- user_id：ユーザの id
- time：検索が行われた時間
- description：Gfdexplorer 上の Query Conditions の部分に表示する絞込み条件
- queryset_id：クエリセットの id

- conditions：絞込み条件の詳細
- count：絞込みの結果，適合したデータの数
- cashed：キャッシュされているかどうか（されていれば 1，そうでなければ 0）

図 11 の表を上から順に見ていく．id1 の検索は，ユーザ id が 1 のユーザが ”キーワード属性 long_name の属性値が Dew point temperature ”という条件でキーワード検索を行い，その結果 15 種類のデータを得たということを示している．この絞込みを id が 1 のクエリセットの中の検索とする．また，この回の検索に行われたのはこの絞込みだけなので，クエリセットの id が 1 である絞込みはこれのみである．検索 id2 の検索は，ユーザ id が 2 のユーザが空間属性においての絞込みを行ったものである．この検索のクエリセットの id は 2 であり，先ほどと同様この絞込みのみで検索が終わっている．そして id が 3〜4 の絞込みは全てキーワード検索である．これらはクエリセットの id が全て 3 になっているので，一回の検索の際に各絞込みの conditions 3 つ全てを条件として絞込みをしたことになる．

そして前述のように，同じ検索中の絞込み，つまり行ったばかりの絞込みとクエリセットの id 同士が同じ絞込みが，Gfdexplorer 上では Query Conditions の部分に示される．

3.8 キャッシュ機能

前述のように，Gfdexplorer で採用している新しい検索インタラクションモデルでは同じ検索セットの中で使用した絞込み条件を同じクエリセットとしてまとめて考えている．我々は検索の内容をテーブル query_histories に，クエリセットの id のみをキャッシュに保存することとした．

キャッシュを利用する理由として，例えばある条件 1 で絞込みを行ったとする．その時，図 12 における Directory” A”が適合したとする．Gfdnavi では親の属性は子も受け継ぐので，ディレクトリ A を親に持つ Directory や Variable もその条件 1 に適合したということになる．次に条件 2 で絞込みを行い，その結果 Variable” b”が適合したとする．そして条件 1 での絞り込みと条件 2 での絞り込みの AND を取ることを考える．この場合，本来は Variable” b”が結果として導かれるはずである．しかし Gfdnavi では子は親の属性も受け継ぐという前提があり，子には親の持つ属性は付与されていないので，検索をするたびにパスの親子関係を確認しなければならず，絞込みに使用する条件が増えれば増えるほど，検索に時間がかかるようになってしまう．これを解消するため，一度行った検索はキャッシュに残しておくことにした．

データベース内に cash というテーブルを用意する．今，図 11 中の id が 5 の検索を行ったところだとする．この検索のクエリセット id は 3 なので，cash テーブルに保存されているクエリセット id が 3 のもの（この場合 id が 3 の検索と id が 4 の検索）との AND を取り，それを検索の結果とする．

しかし現在キャッシュ機能を使うにあたって，キャッシュデータがいたずらに増えてしまう可能性があるという問題点がある．この問題に対して二つの解決策を考案中である．まず一つ目は，絞込み条件が多過ぎる場合に対してである．検索を行う際に使用する絞込み条件の数とクエリセットの数に上限を設けることを考えている．クエリヒストリはずっと残っているので，キャッシュにクエリセットが残っていなくても検索を再現することは可能である．

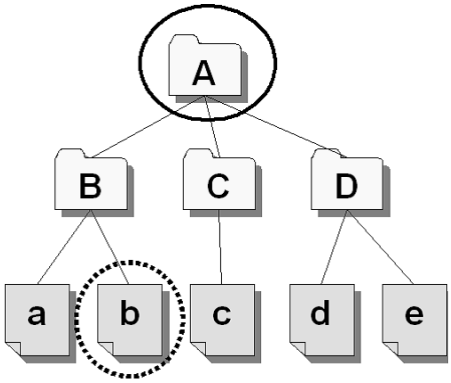


図 12　パスの親子関係

二つ目は，絞込みを行った際に適合するデータがあまりに多いとその検索はあまり意味を持たないので，適合するデータが全体に対して設定されたパーセンテージ以下の場合のみ，キャッシュに保存するという案である．図 11 の id2 の検索は適合するデータが多すぎるのでキャッシュに保存はせず，cashed の値は 0 となる．

また，ディレクトリ構造を考慮したキャッシュの圧縮方法についても考察中である．

4. Web サービス

現在 Gfdnavi は，Web ブラウザ上でインタフェースのインタラクションを操作してデータを検索・公開することができる．このような検索を，ユーザが作成したプログラムの中で実行したい場合も考えられる．例えば，ユーザが実装した分析プログラムを 1990 年から 2006 年までの毎年 7 月南極上空の全データに対して適用したい場合，データの条件で該当ファイル一覧を取得する関数が必要となる．

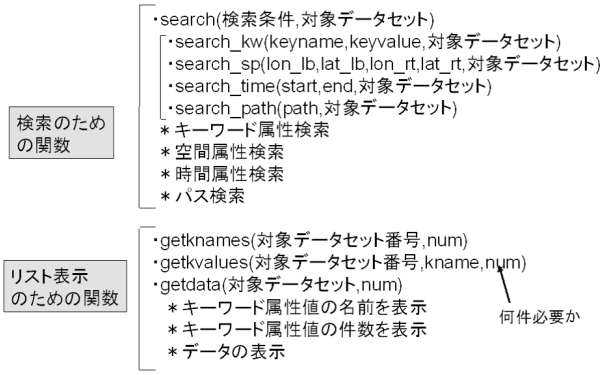


図 13　Web サービスで利用する予定の関数

Web サービス化をすると，絞り込みの条件発行・結果の送受信を XML 形式のプロトコルで行うため，異なるプラットフォーム上のアプリケーションとも統合することが可能になる．つまり Web サービス化の実現により，上記のプログラムに対して，Gfdnavi から取得したデータと気象庁のデーや NASA から取得したデータの両方を使うことも可能である．(主要なデー

タアーカイブサーバでも Web サービスを既に提供している.）

我々は検索インタラクションの処理のために実装し関数をそのまま Web サービスでも利用できるようにする．その関数の詳細を図 13 に示す．今後，実現に向けて考察・実装を進めていく予定である．

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データ属性の探索的検索を利用した，より簡単に低コストでユーザがデータの検索を行うことのできるインタフェースの実装について述べ，更に Web サービス化について提案した．今後は提案した Web サービス化のための具体的な案を考え，実装を進めていく予定である．

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 (課題番号 19024039) 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いただいた森川靖大，神代剛，石渡正樹，林洋介，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] 柳平有美, 渡辺知恵美, 堀之内武:
“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 ,” DBSJ Letters vol.6 No.1

[2] 堀之内武, 西澤誠也, 渡辺知恵美, 森川靖大, 神代剛, 石渡正樹, 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兼デスクトップツール Gfdnavi の開発 ,” DEWS2007

[3] 地球流体電脳倶楽部:
“http://www.gfd-dennou.org/,”

[4] Flamenco Search: UC Berkeley School of Information The Flamenco Search Interface Project
“ http://flamenco.berkeley.edu/index.html,”

[5] mSpace: the School of Electronics and Computer Science, University of Southampton
“ http://www.mspace.fm/,”

[6] NASA Gooddard EOS Data Service:
“ http://eospso.gsfc.nasa.gov/eoshomepage/data_service.php,”

[7] NOAA Statellite and Information Service:
“ http://www.unidata.ucar.edu/software/netcdf,”