汎用データフォーマットを利用した 自然科学データアーカイブスシステムの開発

Development of Science Data Archives System using General-Purpose Data Format

高田　良宏†，笠原　禎也†，尾崎　友紀‡
Yoshihiro TAKATA, Yoshiya KASAHARA, Yuki OZAKI

yosihiro@kenroku.kanazawa-u.ac.jp, kasahara@is.t.kanazawa-u.ac.jp, y-ozaki@cie.is.t.kanazawa-u.ac.jp

† 金沢大学総合メディア基盤センター
Information Media Center of Kanazawa University
‡ 金沢大学工学部
Faculty of Engineering, Kanazawa University
〒920-1192 石川県金沢市角間町
Kakuma-machi,Kanazawa,Ishikawa 920-1192,Japan

概要

大学は研究・教育活動において生み出された情報を蓄積するだけでなく，学外に向けて公開することを求められている．昨今，研究論文，研究資料，教材などの情報を登録し，機関リポジトリとして公開している大学も少なくない．一方，大学には，自然科学分野の実験観測データが大量に蓄積されている．これらのデータは，学術的に非常に貴重であり，学内外から多数の参照要請があるにもかかわらず，データの性質上，それを容易に利活用できる形式での公開が遅れているというのが現状である．そこで，自然科学データを容易に利活用できる形での公開を目的として，自然科学データの中でも規模の大きな地球環境観測データを用い，自然科学データアーカイブスシステムを開発した．開発にあたっては，データの保管，検索，配信，および，利用までを一つの流れとして捕らえて総合的に検討し，蓄積されているデータが有効かつ効率的に利用できるシステムとすることを目指した．本稿では，地球環境観測データの相互利用環境モデルの提案から，システムの設計および実装について報告する．

キーワード：
相互利用環境，データ配信，Web サービス，汎用データフォーマット，地球環境観測

1　はじめに

大学の研究室では日々研究が行われており，多くの情報が生産されている．そして，学内で生産されたさまざまな研究成果，研究資料を電子的な形態で集中的に蓄積･管理し，学内外に公開することを目的として，機関リポジトリ（学術リポジトリ）の構築が急がれている．既に，研究論文，報告書，教材などの登録が行なわれ，公開されている大学も少なくない[1]．一方，自然科学分野の実験観測データは，学術的に非常に貴重であり，学外から多数の参照要請があるにもかかわらず，データの性質上，それを容易に利活用できる形式での公開（提供）が遅れているというのが現状である（2 章参照）．

このような背景のもと，筆者らは，大学内の学術情報を一元的に管理し，情報を提供する総合実験データベース[2,3]を構築中であるが，その際に問題となるのが自然科学系の実験観測データである．これらのデータは，一般に，容量が膨大でデータフォーマットが非常に複雑かつ多様なためである．

本研究では，自然科学データの管理，提供の指針となるべく，自然科学データの中でも規模の大きな地球環境観測データを対象として，相互利用環境モデルを提案し，自然科学データアーカイブスシステムを開発した．開発にあたっては，データの保管，検索，配信，

および，利用までを一つの流れとして捕らえて総合的に検討した．さらに，開発したシステムが，自然科学データ全般に対し汎用的に利用できることを目指した（2.4 節参照）．なお，今回用いたデータは，本学で蓄積・管理している「あけぼの衛星」による地球周辺の電波環境観測に関するデータである．

本論文の構成は次の通りである．まず，2 章では地球環境観測データの現状と問題点を整理し，続いて，3 章で今回提案した相互利用環境モデルの概要を述べる．さらに，4 章で設計の要点，5 章で実装について述べる．最後に 6 章でまとめを行なう．

2　地球環境観測データの現状と問題点

本研究で主に取り扱う地球環境観測データは，自然現象を対象としているため，二度と再現することができない希少データであり，学術的に非常に貴重なものである．しかし，これらは，十分に活用されているとはいえず，結果的に死蔵されているケースが少なくない．本章では，地球環境観測データの現状と問題点を整理する．

2.1　地球環境観測データ

地球環境観測は，気象，海洋，地震，電離層，磁気圏など多くの分野で行なわれ，地球環境の研究が進められている．地球環境分野の観測は，地上，地中，水中などに観測器を設置して行う観測から，船上，航空機上，ロケット，科学衛星上に観測器を搭載する方法まである．非常に多種多様な観測手段を用いるため，データの特性やフォーマット，観測の分解能もそれぞれ異なる．さらに，観測技術，伝送技術の進歩，大容量ストレージの登場などによって，詳細なデータを得られるようになり，一つのプロジェクトで容量が十数 TByte を超えることは珍しいことではない．

2.2　管理の現状

地球環境観測データは，近年，各国で集中管理・公開するためのデータベース化が進んでいる．米国では，米国航空宇宙局(NASA)の国立宇宙科学データセンター(NSSDC)や米国海洋大気庁(NOAA)などで観測データを集中管理している．それに対し，日本では，気象庁や宇宙航空研究開発機構(JAXA)などで一部の地球環境観測のデータを蓄積・公開しているものの，全ての観測データをカバーするに至っておらず，多くの貴重な観測データは，大学などの研究室に分散して蓄積されている．これらの地球環境観測データの集中管理が困難なためであるが，その大きな要因として二つ挙げることができる．

一つ目の要因は，地球環境観測データの観測に関わった研究室に，それらデータの管理・公開の責任があることである．特に，日本では，プロジェクト担当の研究者がデータの解析だけではなく，観測機器の開発や実際の観測をはじめ，データの蓄積，管理，公開まで責任を持つ場合が非常に多い．

二つ目の要因は，観測データの特性やフォーマットが非常に多種多様かつ複雑なことにある．フォーマットは各分野，各プロジェクト，さらには，観測機器毎に異なる場合も多い．これらは独自フォーマットといわれる．また，観測データは一般にバイナリ形式で大規模なものが多い．このようなバイナリ形式の独自フォーマットは，プラットフォーム間での互換性が低く，データのフォーマットを熟知しないとデータを読み出すことすらできない．

2.3　相互利用の現状

大学の研究室などに分散して蓄積されている地球環境観測データがバイナリ形式の独自フォーマットであることは，相互利用を行なう際に非常に問題となる．提供側の研究者は，データの提供時に，データの構造，データの意味，利用法，サンプルプログラムとその解説をマニュアルとして準備し，データと共に渡す必要がある．被提供者側の研究者は，それらを基に独自に低レベルの I/O 操作プログラムや解析ソフトを開発しなければならない．さらに，データの相互利用のための効率的な配信手段が確立されていないという問題もあり，その方法は，メールに添付，ftp や，http などを利用したダウンロード，記録媒体を郵送するなどさまざまである．

現状では，相互利用を行なおうとすると，観測データ毎に個別の対応が必要であり，研究者は本来のデータ解析業務以外に，様々な労力を割く必要がある．これらのことは，データの提供者，被提供者にとって重い負担であり，相互利用の妨げとなっている．

データフォーマットの問題に関しては，標準化され互換性が高い記述法として，IT 分野で積極的に利用されている XML の適用が考えられる．過去に著者らは，メタデータの配信に XML を適用した[4]. XML はメタデータの配信には有効である．しかし，今回取り扱うバイナリ形式の地球環境観測データにテキストベースの XML を適用すると，容量が元データの数倍から数十倍となり，保存，配信（通信）の観点からみて現実的でない．

2.4　汎用的な配信システムの必要性

地球規模の全体像を研究し，地球環境を理解するには，それらの観測データを相互参照し，総合的に検討する必要がある．今後，従来型の単一種の観測データに着目した研究から，観測機器，各プロジェクト，さらには，分野を超えた複数種の観測データの相互比較を行なう研究スタイルへの移行は必須である．

このため，相互利用の妨げとなっている配信手段および独自フォーマットの問題，提供者，被提供者にかかる負担問題などをクリアし，個々の観測データが容易に利活用できる形式で公開（提供）可能なシステムの開発が必須である．しかし，このような，分野を超えたデータ相互利用に対する汎用的な配信システムは未だに実現されていない．

3　　自然科学データアーカイブスシステム

3.1　相互利用環境モデル

分散して蓄積・管理されている実験観測データを相互利用するための自然科学データアーカイブスシステムを開発するにあたり，まず，その基礎となるモデルを提案する．図 1 にその概要を示す．

このシステムでは，汎用的なフォーマットを適用した実験観測データならびにそのメタデータを配信するための配信システムを構築し，実験観測データを蓄積・管理している研究室に分散的に配置する．環境の構成に集中管理センタ（サーバ）は存在せず，各配信システムは独立している．データ管理者は，各自が保有する実験観測データおよびメタデータのみを管理する．クライアントも，集中管理センタ（サーバ）を経由せず直接配信サーバにアクセスする．結果的に各地に分散したデータが仮想的に統合された環境の構築が可能となる．メタデータ（配信サーバ）の所在は，現在各地で構築が行なわれている機関リポジトリで集約的に管理する．

本研究では，上記モデルの実現のために，開発の指針となる諸元の提示，および，システムの基本構成の設計，実装を行なう．なお，システムの基本構成は実験観測データおよびメタデータ配信システムとクライアント（図 1 中の一点鎖線で囲まれた部分）である．

3.2　諸元

3.2.1　配信環境

システムがデータ配信に利用する方式は，インターネット上で利用される一般的なプロトコルの使用が望ましい．ただし，昨今インターネットを介した攻撃が激増し，セキュリティ向上対策の一つとしてファイアウォールを設置するのが常識であり，その通過を考えた場合，特別なプロトコルを許可しない機関があると考えられ，相互利用に支障をきたす可能性が高い．そこで，配信に使用する通信プロトコルは，手続きの一般性，ファイアウォール対策，保守性を考慮し，http, https または smtp などの利用が望ましいと考えられる．

3.2.2　配信データのフォーマット

実験観測データの配信にあたっては，いくつかの分野で規格化され，使用されている汎用データフォーマットの導入を検討する．独自フォーマットで保存され

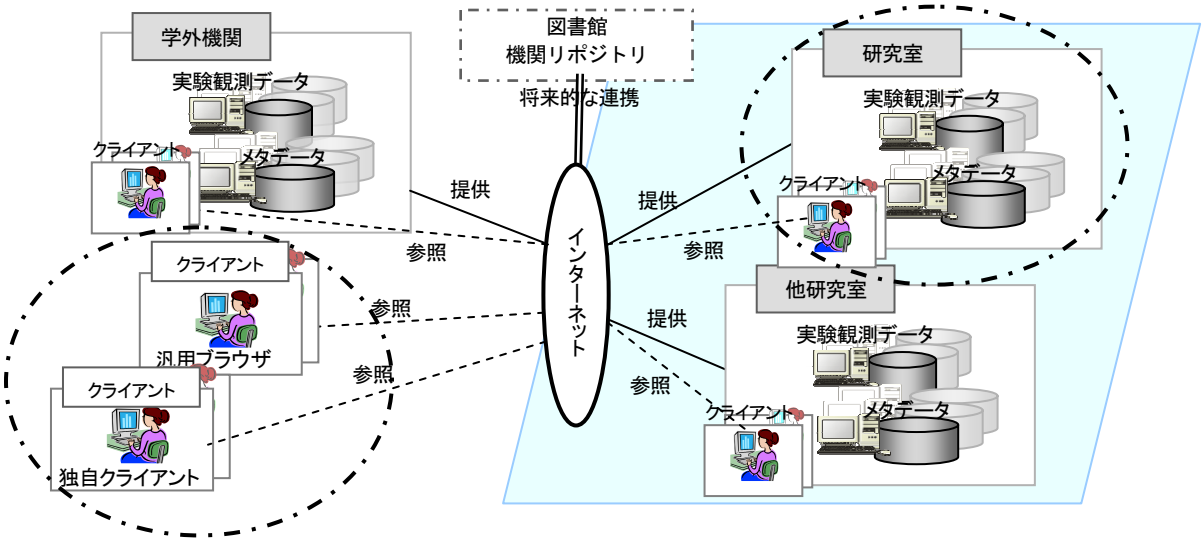


図 1　自然科学データの相互利用環境モデル

ているデータに対して，その分野の背景や観測データの構造の類似性を検討し，最適な汎用データフォーマットを提案し，データ配信に利用することとする（4.2 節参照）．

3.2.3 クライアント

クライアントは，多様なプラットフォーム・OS に対応できることとし，管理者からの提供版とユーザ作成版の二本立てとする（4.3 節参照）．また，ユーザ作成版は，ユーザがそれぞれの利用形態に応じたものを容易に作成可能なように配慮する．

3.2.4 サーバ環境

配信システムは，実験観測データを蓄積・管理している一般的な研究室レベルでの運用となることから，初期コスト，維持費などを考慮し，PC を使ったサーバ，Linux，オープンソースで構築を行なう．

4 システムの設計と基礎実験

３章で検討した開発諸元に基づき設計を行なった．本章では，設計においての要点を説明する．

4.1 配信環境

4.1.1 Web サービス

実験観測データの検索（メタデータの配信），実験観測データの配信は，それぞれインターネット上の汎用的なサービスとして提供する．そのため，利用する配信手法は次に示す条件をクリアすることが必要である．

- 配信するデータに対して柔軟性が高いこと（データの種類，データの規模など）．バイナリ型のデータへの対応は必須である
- サーバの機種，OS，実装言語に依存しないこと
- ネットワーク上で独立したサービスを提供できること（他のサーバに依存しないこと）
- 配信処理の自動化に容易に対応できること
- 保守性，可用性などにも優れていること

配信手法の決定に際しては，まず，OS・ベンダに依存する手法やトランスポートプロトコルに依存する Socket 通信などは汎用性の面から除外した．また，smtp は一般的なプロトコルであるが，大容量のデータ配信を行なうため，各機関のメール系の基盤設備に掛かる負担が大きいと判断し除外した．

最終的に SOAP over HTTP を用いた Web サービス[5,6,7]と著者らが素材管理システム開発で提案した rsync over ssh を用いた予約・配信法[8]について検討した．rsync を用いた予約・配信法は学内において，小容量データの配信では，既に実績を積んでおり，安定したサービスを提供している．しかし，本システムで必要とするネットワーク上で独立したサービスを提供できることという点，サーバが学外に分散している点，対象とするデータの規模の問題など，解決しなければならない問題がある．一方，Web サービスは，実績は少ないが上記の全ての条件をクリアしている．そこで，いくつかの既存システムを Web サービスに置き換え，動作確認を行なった．図 2 は画像簡易検索システムに適用した事例の概要である．さらに，データ規模と配信速度に関して配信実験を行なった（4.1.3 節参照）．結果として，十分に機能し，さらに，研究室の分散，データの種類やデータ規模の問題もクリアできると判断し SOAP over HTTP を用いた Web サービスを本システムの配信方式に採用することとした．

4.1.2 サーバの分離

サーバには，図 3(a)に示すように，データ配信機能に特化したデータサーバと，Web ブラウザからのアクセスを可能とするための Web サーバがある．また，データサーバには提供用の実験観測データを配信する機能のものと，それらのメタデータを配信するものがある．これらは，同一マシン上での実現も可能であるが，小規模 PC サーバでの実現，拡張性，保守性を考慮して分離する．

サーバの配置を図 3（b）に示す．データサーバは実験観測データの内容を熟知し，新規登録などを責任もって行える研究室で管理されるべきであるので，研究室で管理する（①）．ただし，Web サーバは管理者の負担を考慮しセンターなどで集中管理することも可能とする（②）．

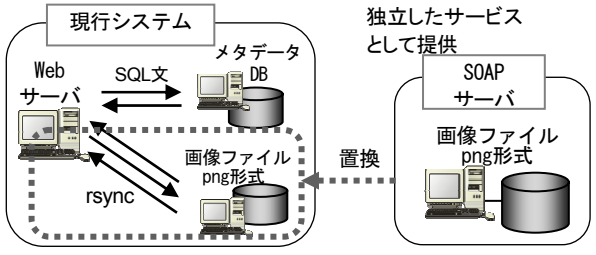
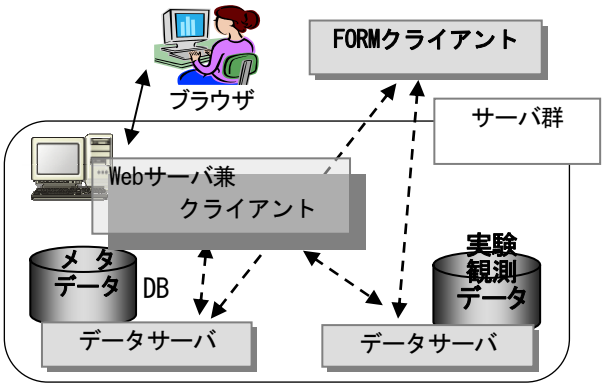
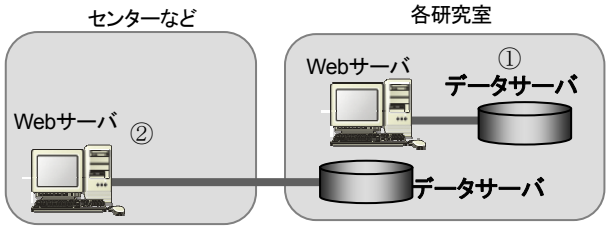


図 2 Web サービスに置換した事例



(a):構成概要



(b):サーバの配置

図 3　サーバの分離

4.1.3　配信実験

実験用システムを構築し，テストデータを使用して配信実験を行なった．配信速度は，配信システムを評価する上で，大きなウエイトを占める指標である．また，サーバの選定など，実行環境の整備にも影響する．

図 4 に配信実験の概要およびスペックを示す．データサーバ上には配信サービスが動作しており，クライアントからの要求により，データの転送を開始する仕組みである．実験観測データを想定し作成した 1MByte, 10MByte，100MByte，1GByte のバイナリデータの配信時間などを測定した．測定は，(1)-(2)配信サービス呼び出し準備時間，(2)-(3)配信時間（データ転送要求～転送終了），(1)-(3)全体時間で行なった．データの配信は，SOAP メッセージにデータを添付する形式（マルチパート MIME）で行なった．表 1 は実験観測データを想定し作成した 1MByte，10MByte，100MByte，1GByte のバイナリデータを配信したときの平均配信時間を示す．

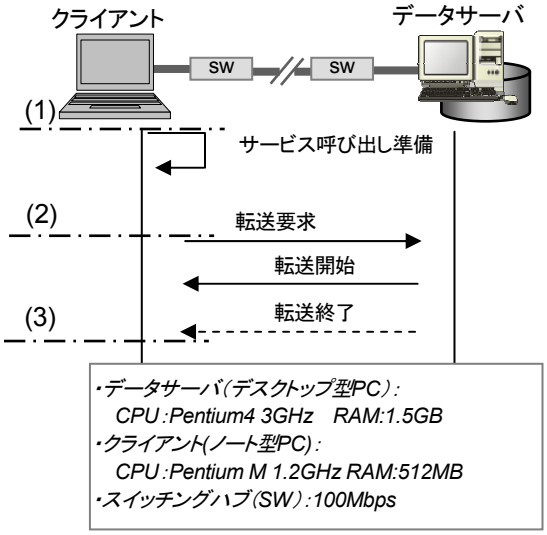


図 4　配信実験環境

表 1　バイナリデータの平均配信時間

単位 Byte→	1M	10M	100M	1G
(1)-(2)(ms) 準備時間	1,408	1,428	1,411	1,410
(2)-(3)(ms) 配信時間	480	1,392	13,790	139,057
(1)-(3)(ms) 全体時間	1,888	2,820	15,231	140,467
速度（Mbps） 配信時間 全体時間	16.7 4.2	57.5 28.4	58.0 52.5	58.9 58.3

配信サービスを呼び出すための準備時間は平均で 1.4 秒であった. この値はサービスの規模によって多少増減はあるが, 配信するデータの大きさに依存しない. 配信時間に対する速度は, 10MByte～1GByte のデータでは 50Mbps 以上の値となり，十分高速であるといえる．1MByte の速度が約 16.7Mbps となった．これは，小さなファイルでは転送要求処理と終了処理の影響が大きいためである．一方，運用を考えた場合は全体時間に対する速度が重要であるが，1MByte，10MByte，100MByte，1GByte のデータの配信速度は，それぞれ，4.2Mbps，28.4Mbps，52.5Mbps，58.3Mbps と実用的な値となった．

4.2　汎用データフォーマット

4.2.1　汎用フォーマットの条件

先述（2.2 節参照）の通り，地球環境観測データを独自フォーマット形式で提供することは現実的ではないので，いくつかの分野で規格化され，使用されている汎用的なデータフォーマットを適用することとした．提供用の汎用フォーマットに求められる条件を表 2 に示す．

4.2.2　調査

現在，地球環境観測分野で広く使われているフォーマットを調査した. それぞれのフォーマットについて，

表 2　汎用データフォーマットの条件

1. 自己記述型:データファイルを読めばデータが理解できるように,ファイルに変数やデータなどに関する属性,意味,説明などの自己情報を含めることができること
2. バイナリデータをサポート:観測データは非常に大規模なものが多い.そのため観測データでは一般的なバイナリ形式をサポートすること.場合によってはデータ圧縮も可能なこと
3. 分野に特化しない:特定の分野に特化せず,広くデータ交換用として利用できること
4. 機種などに依存しない:バイナリデータの格納方式はOSやプラットフォーム（ハードウェアアーキテクチャ）により異なる場合があるが,格納方式が異なるコンピュータ間でもデータを共有できること
5. 汎用ソフト:解析ソフト作成の負担軽減のため,市販,フリーの汎用ソフトでも解析できること
6. ランダムアクセス:容易なデータアクセス環境を提供するためにランダムアクセスが可能であること

汎用データフォーマットとしての条件に適合しているかを評価した．表 3 に評価の結果を示す．今回は自己記述性と，適用可能な分野の広さを重視した．自己記述型フォーマットとはデータファイルに自己情報（ファイルの属性，変数の属性，次元，データの説明，単位，グラフ座標，データ利用規定，参考文献，コメントなど）を保存することが可能で，ファイルを読めばデータを理解できるファイル構成を実現できる．netCDF[9]，CDF[10,11]，HDF[12]は，自己記述型の汎用データフォーマットとしての適性を備えており，また，自然科学分野全般のデータに適用可能である．この 3 つのフォーマットを配信用のフォーマットとして推奨することとする．なお，3 章でも述べたが，独自フォーマットで保存されているデータに対して，その分野の背景や観測データの構造の類似性を検討し，適切と思われる汎用データフォーマットを選択するのであって，天文分野や月・惑星系衛星観測などで使用されている FITS，PDS などを否定するものではない．

4.2.3　事例

テストデータを netCDF，CDF について適用し，実装の容易さ，操作性，汎用解析ツールなどとの親和性などの確認を行なった．作成時のサポートとして，両フォーマットとも人間指向型設計機能を持つ．記憶形式はバイナリであるが，ファイルの作成（設計）時には，設計図とも言える人間指向型のテキストファイルを作成し,それを基にファイルを作成する機能を持つ．

表 3　地球環境観測関係のデータフォーマット

	自己記述性	用途 主な利用分野	自然科学分野全般での利用
netCDF	○	**科学データ交換用** 気象，海洋他	**可**
CDF	○	**科学データ交換用** 太陽地球系物理他	**可**
HDF	○	**科学データ交換用** 地球環境観測衛星他	**可**
FITS	○	天文データに特化 天文学	不可
PDS	△	天体データに特化 月・惑星系衛星観測	不可
GRIB	△	天気情報に特化 気象，天気	不可

逆に，ファイルから人間指向型のテキストファイルを生成する機能も持つ（図 5）．データへのアクセスに関しては，機種に依存しない（表２の４項目目参照）ライブラリが充実しており，低レベルの I/O 操作を意識する必要がない．

netCDF の人間指向型ファイル（CDL）を用いた表記例を図 6 に示す．これは説明のために簡単化したものであるが，ファイル中に，ファイル属性，変数名，候補最小／最大値，グラフ化される場合の座標，データおよび注釈が記述されている．図 7 は図 6 の内容で生成した netCDF を汎用解析ツール（IDL）によりテスト描画したものである．描画に際しては，利用者が IDL にデータ属性や座標に関する情報を与えなくても，netCDF 自身が持つ情報のみ使用して描画される．

なお，CDF については，5.2.2 節において後述する．

4.3　クライアント

管理者からの提供版は，検索，表示，保存などの基本的機能を搭載した Web サーバ版と Windows, Linux,

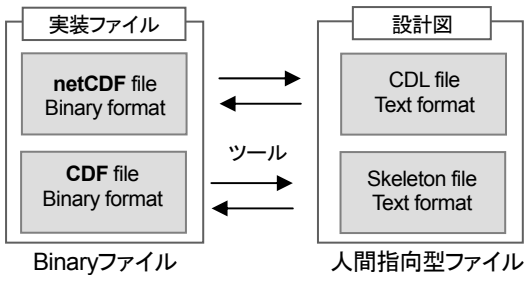


図 5　作成時のサポート

```
netCDF sample {
      dimensions:     // 次元変数
            row = 256 , column = 256 ;
      variables:          //   変数
            // 主変数
            short   value ( row , column ) ;
                  value : valid_min = 1 ;
                  value : valid_max = 256 ;
            // 座標変数
            short row( row ) ;
            short column( column ) ;
            // グローバル変数
            :title = “Sample data” ;
            :history = “20060130” ;
       data:
                  row =   1,2,3, ・・・ ,255,256 ;
                  column =   1,2, 3, ・・・ ,255, 256 ;
                  value =   1,2,3,   ・・・ ,255, 256 ;
}
```

図 6　表記例

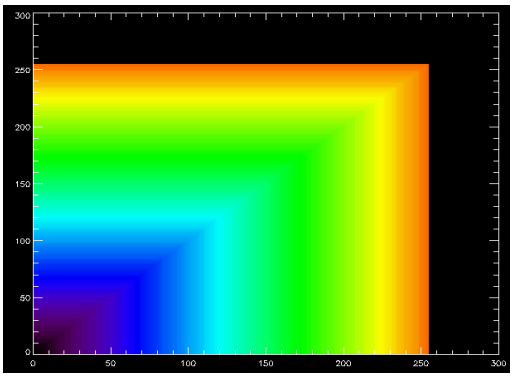


図 7　汎用解析ツール(IDL)によるテスト描画

Mac などに対応するアプリケーションとする．なお，Web サーバ版とは，ブラウザからの利用を可能とするために，Web サーバ上に Web サービスのクライアント機能を組み込み，Web アプリケーションとして提供するものである．作成言語は Web 関係のアプリケーションの作成に実績がある Java および JSP とする．

ユーザ作成版は，研究者が自分の研究スタイルに応じて作成するクライアントであり，例えば，独自検索画面，自動配信（受信），解析システム組み込みといった多様な形態が考えられる．ユーザ作成版に関しては，ユーザがクライアントを作成する場合の支援として，クライアントがサーバに接続し通信を行なうための一連の手続きを実装したモジュールを提供する．

5　実装

前章での設計を基に，自然科学データアーカイブスシステムの実装を行なった．同時に，実際の実験観測データを用いて保存用，公開用のファイルの作成を進めている．本章では，実装したシステムの概要と，すでに汎用データフォーマットの適用が終了した実験観測データについての導入事例を紹介する．

5.1　配信システムの構築

5.1.1　概要

ここでは，自然科学データアーカイブスシステムの基本構成（3 章参照）である実験観測データ配信システムとメタデータ配信システムおよびクライアントを実装した．図 8 に Web サービス技術（SOAP over HTTP）を用いて実装した実験観測データ配信システムおよびメタデータ配信システムの概要を示す．

SOAP サーバは OS が Linux，システムの構築は Apache Axis[13] + Java で行なった．ハードウェアはデスクトップ型 PC（CPU : Pentium4 3GHz ，RAM:1.5GB）を使用した．また，メタデータを蓄積している DB サーバも SOAP サーバと同等のスペックである．

5.1.2　サービスとメソッド

表 4 に実験観測データ配信用およびメタデータ配信用の各サーバ上で動作するサービス（プログラム）の概要を示す．メタデータ提供用サービスは，観測日付よりファイル名などのメタデータを検索し，該当するメタデータを配信するものである．実験観測データ提供用サービスはファイル名を受け取り，該当する観測データを配信するものである．

また，メタデータ提供用サービスは SOAP-RPC 方式，実験観測データ提供用サービスは SOAP-RPC+SOAP Messages with Attachments[14]方式でデータ配信を行なっている．

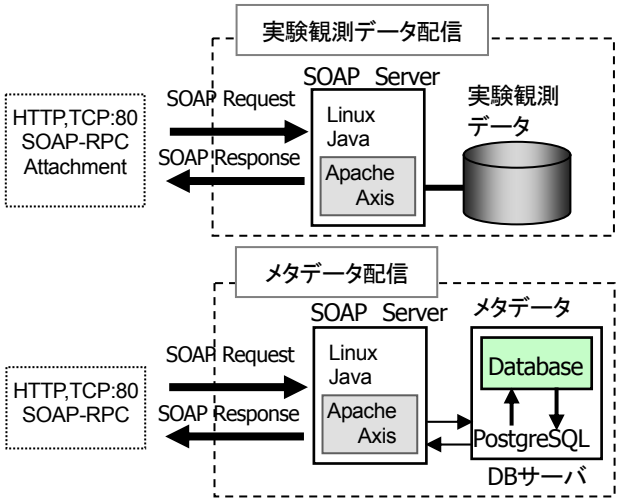


図 8　配信システム

表 4　サービスとメソッド

機能	サービス名/メソッド名
メタデータ提供用	**xxxyyy**SearchService **xxxyyy**Search
実験観測データ提供用	ScienceArchivesService get**xxx**File

xxx：データの種類　**yyy**：フォーマットの種類

5.1.3　クライアント

先述の通り，クライアントは，データ管理者が作成し，配布するものと，研究者がその利用形態に応じて作成するものがある．管理者が提供する基本的な機能を備えたアプリケーション版と Web サーバ版を作成した．図 9 はアプリケーション版クライアントのスナップショットである．

5.2　汎用データフォーマットの導入例

5.2.1　データの説明

利用するデータは「あけぼの衛星」で観測した地球周辺電波に関するデータで，デジタルデータが 1.5TByte 以上，元データがアナログ形式であるものをデジタル化した 15TByte 以上にのぼる大容量データである[15,16]．今回はその中の，マルチチャンネル解析装置(MCA)で測定されたデータを利用した．

一般に科学衛星観測におけるデータの伝送条件は非常に厳しく，「あけぼの衛星」では最速モードで 64kbps（他に 16kbps/4kbps モードあり）に留まる．このため，限られた帯域において，できるだけ多くの観測データを伝送できるように，データは全てバイナリ形式で，そのフォーマットは，使いやすさを犠牲にした複雑なものとなっている．また，過酷な条件下での観測，伝送を行なうため，他観測機器や自然界からのノイズなどの影響によりデータに欠測やエラーが混在する．これらの特徴を整理すると次の通りであり，汎用データフォーマットの導入例（評価）としては最適な事例である．

- 容量が膨大である
- バイナリ形式のデータである
- フォーマットが非常に複雑である
- 欠測値やエラーデータが多数混在している

5.2.2　CDF

今回「あけぼの衛星」の MCA データに適用した汎用データフォーマットは CDF（Common Data Format）である．CDF は米国宇宙計画の一部として NASA により開発された自己記述（自己表現）形式のデータフォーマットで，非常に汎用性が高い．地球周辺の環境を観測する科学衛星関係で利用されている他，自然科学データの交換に広く使用可能であり，日本規格協会（JSA）の IEC 活動推進会議でも科学データ交換標準フォーマットとして紹介されている[17]．ファイル中にデータと共に，属性（ファイル，変数，データ），データの扱い方，注意事項などが記憶できる．

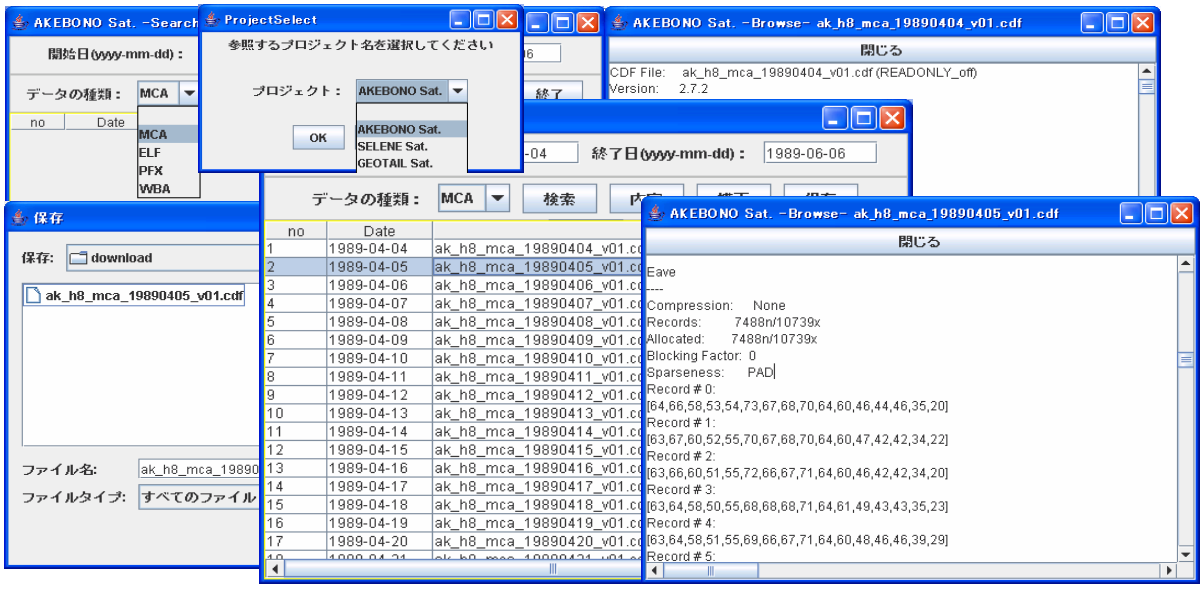


図 9　クライアントの実装例（アプリケーション版）

5.2.3　フォーマットの設計方針

設計の大前提は，CDF の自己記述性を活かしファイルを読めばデータを理解できるファイルとすることであり，その実現のために十分に時間を掛けて様々な角度から検討を行なうことである．そのため，フォーマットの設計にはコストが発生するが，一旦適切な設計を行なえば，以降，作成はデータを機械的にフォーマットに適用するだけでよい．

また，設計時にファイルの属性，変数の属性，次元など必須項目以外に，次の点に注意する必要がある．

- 解析・可視化を想定し，データの単位，想定最大値・最小値，グラフ軸（スケール，ラベル），表示（印字）時のフォーマットなども定義する
- 注釈（コメント）は詳細に記述する．例：ファイル・変数・データの意味，観測条件，較正法・バージョン，利用条件，責任者，関連文献，etc.
- 各分野で既にフォーマットの詳細が定義されている場合はそれを優先する

実際の設計では，長期保存用と公開用の 2 種類の設計を行なった．長期保存用とは，研究室での保存用として，観測データ（生データ）の内容に手を加えず，そのまま CDF 化するものである．これは，必ずしも本論文で述べる公開システムに必要なものではないが，将来的に，より高次のデータを生成する必要が発生したときに，生データに戻る必要がないため，結果的にデータの再利用性の向上につながると考えられる．また，観測データは，較正処理を行なわないと物理量としての意味を持たないため，公開用ファイルは，利用者がそのまま利用できるように，データ較正を施したものを CDF 化する．設計を適切に行なっておけば，公開用ファイルの作成は，長期保存用ファイルから簡単に行なうことが可能である．このように，長期保存用ファイルも公開用ファイルも CDF 化することにより一元的な管理が可能となる．

5.2.4　CDF 作成

長期保存用ファイルと公開用ファイルを作成した．作成したファイルへのアクセスであるが，専用のプログラムを作成しなくても，CDF の開発者から提供されている汎用アクセスプログラムや市販の汎用解析ソフト（IDL，MATLAB など）を用いて簡単に行なうことができる．これらは，ファイル中に収められた自己記述情報を利用してデータにアクセスしている．図 10 は，汎用アクセスプログラムを使用して，公開用の CDF を読み込み，整形して画面に表示した例である．同様に図 11 は汎用解析ソフトの MATLAB により可視化した例である．

```
File Info                                              ファイルの情報
==========
CDF File:       ak_h8_mca_19890405_v01.cdf (READONLY_off)
Version:        2.7.2
Format:         SINGLE
 ・・・省略・・・
Global Attributes (22 attributes)                      データ全体に係わる属性
==========
Acknowledgement (1 entry):0 (CDF_CHAR/119): AKEBONO VLF-MCA ...
Descriptor (1 entry): 0 (CDF_CHAR/28): MCA > multi channel analyzer
PI_affiliation (1 entry):0 (CDF_CHAR/19): KANAZAWA University
 ・・・省略・・・
Variable Attributes (17 attributes)                    変数に係わる属性
==========
VAR_TYPE:    Emax (CDF_CHAR/4): data
             Eave (CDF_CHAR/4): data
             Bmax (CDF_CHAR/4): data
             Bave (CDF_CHAR/4): data
             PostGap (CDF_CHAR/4): data
             Epoch (CDF_CHAR/12): support_data
             channel (CDF_CHAR/12): support_data
UNITS:       Emax (CDF_CHAR/2): dB
 ・・・省略・・・
Variable Information (7 variables)                     変数に係わる情報
==========
Emax       CDF_UINT1/1   1:[16]   T/T sparseRecord="PAD"
Eave       CDF_UINT1/1   1:[16]   T/T sparseRecord="PAD"
Bmax       CDF_UINT1/1   1:[16]   T/T sparseRecord="PAD"
Bave       CDF_UINT1/1   1:[16]   T/T sparseRecord="PAD"
PostGap    CDF_UINT4/1   0:[]     T/ sparseRecord="PAD"
Epoch      CDF_EPOCH/1   0:[]     T/ sparseRecord="PAD"
channel    CDF_FLOAT/1   1:[16]   F/T sparseRecord="PAD"
 ・・・省略・・・
Variable Data (7 variables)                            データおよび付加情報
==========
Emax
----
Compression:       None
Records:           7488n/10739x
Allocated:         7488n/10739x
Blocking Factor:   0
Sparseness:        PAD

Record # 0:[61,67,58,53,56,83,80,72,74,69,63,50,48,48,42,31]
Record # 1:[61,68,62,53,56,82,79,71,76,69,64,50,46,45,41,32]
Record # 2:[61,67,58,54,58,80,77,71,75,69,66,50,46,45,41,25]
Record # 3:[62,65,57,50,58,78,81,71,75,69,66,52,46,45,42,30]
Record # 4:[62,65,59,53,56,79,71,70,76,68,64,50,50,47,42,36]
 ... 省略..
```

図 10　汎用アクセスプログラムによる内容表示（整形表示）

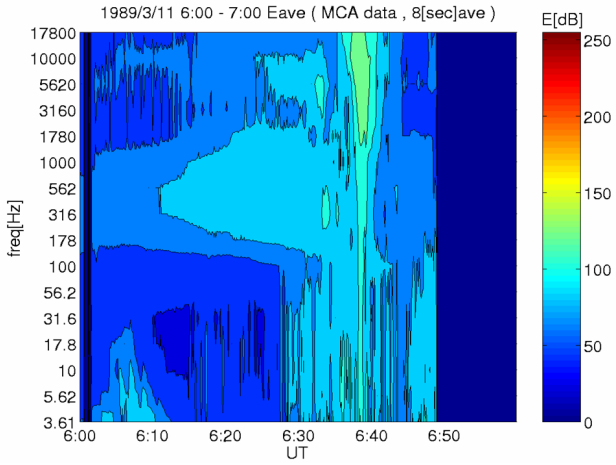


図 11　汎用解析ツール(MATLAB)による可視化例

6　まとめ

本研究では，自然科学データを容易に利活用できる形での公開を目的として，相互利用環境モデルを提案，さらに，その実現のために自然科学データアーカイブスシステムの基本構成の開発を行なった．

地球環境観測データが抱える問題の内，実験観測データが各地の研究室に分散しているという問題に対しては，Web サービス技術を取り入れた配信システムとしたため，各研究室の事情（プラットフォーム，OS，実装言語）をシステムで吸収することが可能となった．配信実験の結果，Web サービスを用いた配信システムが，本研究で取り扱う大容量の実験観測データの配信に十分利用できるという結論を得た．

バイナリ形式の多種多様なフォーマットが存在するという問題に対しては，自己記述型の汎用データフォーマットを適用することにより流通性を高めることができた．特に，netCDF，CDF，HDF などはデータの自己記述性が高く，さらに作成から解析に至るまでの操作性に優れていることを示した．今回，「あけぼの衛星」の MCA データに CDF を適用したが，CDF 化により操作性，互換性に優れたファイル構造となった．また，CDF 化により，長期保存用ファイルから公開用ファイルまで一元的な管理が可能となった．

さらに，上記の手法を組み合わせ，汎用フォーマット化されたデータを SOAP メッセージに添付する方式（マルチパート MIME）としたため，フォーマットの種類やデータの内容に依存しないシステムとなった．これにより，地球環境観測データが抱えるデータの分散，多様なフォーマットという問題を同時にクリアすることが可能であり，分野を超えたデータ相互利用に道を開くものである．

今回開発したシステムは，相互利用環境モデルの基本構成として十分実用的に機能するものである．また，「あけぼの衛星」のデータに依存しない汎用的な仕様であるため，多様なデータに対応可能である．

今後は，機関リポジトリとの連携によるメタデータ（配信サーバ）の所在の集約的管理を含め，継続して検討・改良を行い，自然科学分野の実験観測データが容易に利活用できる研究環境の向上につなげていきたいと考えている．

参考文献

[1] 金沢大学学術リポジトリ，
http://dspace.lib.kanazawa-u.ac.jp:8080/dspace/

[2] 金沢大学総合メディア基盤センター, COM.CLUB, Vol.27, 2003.

[3] 笠原 禎也, 金沢大学における実験データベースの構築, 国立情報学研究所 学術情報ネットワーク(スーパーSINET/SINET) 成果報告集, pp221-228, 2004.

[4] 高田 良宏, 笠原 禎也, 大林 誠, 田中 祥平, 大規模な科学データベースの構築と効率的なデータ検索配信システムの開発, 学術情報処理研究, pp.33-43, No.8, 2004.

[5] W3C Note, Web Services Architecture, http://www.w3.org/TR/ws-arch/

[6] W3C Recommendation, SOAP Version 1.2, http://www.w3.org/TR/soap/

[7] 本 俊也, Web サービス マスタリングハンドブック, 秀和システム, 2004.

[8] 高田 良宏, 笠原 禎也, 佐藤 正英, 鈴木 恒雄, 松本 豊司, 森 祥寛, e-Learning 素材管理・再利用システムの開発, コンピュータ& エデュケーション, Vol.20, pp.68-73, 2006.6.

[9] C 版 netCDF ユーザマニュアル Version3, Unidata Program Center, Russ Rew, Glenn Davis, Steve Emmerson, and Harvey Davies, （日本語訳）地球流体電脳倶楽部, 1999.

[10] CDF 3.0 User's Guide 日本語版, Goddard Space Flight Center, NASA,村田 健史(訳),2005-2006.

[11] CDF 3.0 User's Guide, Goddard Space Flight Center, NASA, 2005.

[12] HDF5 User's Guide Release 1.6.5, NCSA, 2005.

[13] WebServices – Axis , http://ws.apache.org/axis/

[14] W3C Note, SOAP Messages with Attachments, http://www.w3.org/TR/SOAP-attachments/

[15] 昭和 63 年度第 2 次飛翔実験科学衛星 EXOS-D (M-SII-4)計画書, 宇宙科学研究所 SES データセンター, 1989.

[16] I.Kimura, K.Hashimoto, I.Nagano, T.Okada, M.Yamamoto, T.Yoshino, H.Matsumoto, M.Ejiri and K.Hayashi, VLF Observation by the Akebono (EXOS-D) Satellite, 42(4), pp459-478, 1990, J.Geomag.Geoelectr.

[17] 日本規格協会 IEC 活動推進会議, Scientific Data Exchange Standards, http://www.iecapc.jp/06/diffusenew/standards/science.html