土壌・地表面気候データを中心とする地球環境デジタルライブラリの試作

生駒栄司[†,☆] 沖 大幹[††] 喜連川 優[††]

近年の地球環境への関心の高まりとともに, リモートセンシングデータを始めとする多様な地球環境データをデジタルライブラリ化したシステムの開発が多方面で行われつつある. しかし, 利用可能なデータ数およびシステムの操作性の両面で, 地球環境工学分野の研究レベルにおいて十分に利用可能なシステムが存在していないのが現状である. このような背景から我々は土壌・地表面気候に関するデータを対象とし, デジタルライブラリへのデータローディングツールを新たに開発したが, 本論文ではその方式ならびに定量的評価結果について報告する. また, ユーザインターフェースに関しては, 地球環境研究者とともにその要件を検討し, 現行システムでは支援されていない多くの機能を有する検索インターフェースを開発し, 地球環境 Web サイトを構築した. 当該ローディングツールにより, 約 1100 種類 30000 個のファイルをほぼ自動的にダウンロードすることが出来た. また, 土壌・地表面気候関連という地球環境の限られた専門家を対象とするにも拘わらず, 数ヶ月の運用により月間アクセスは 8000 件に達しており, 国内外の専門家にとって有用なシステムを構築することが出来た.

Development of an earth environmental digital library system for soil and atomonpheric data

Eiji Ikoma ,[†,☆] Taikan Oki[††] and Masaru Kitsuregawa[††]

Recently as the interest to the earth environment increases, research on a digital library system is getting much more popular which tries to interpret various kinds of earth environmental data such as remote sensing data. However, no system can be found for practical use at the field of earth environmental research concerning the variety of including data and user interface system. In this paper, we propose new methods about data loading system at the phase of soil and atmospheric data injection and user interface system for searching data. We develop an earth environmental digital library and operate on the Web. Using this method enables to introduce 80% of data automatically for injecting 1100 kinds of earth environmental data. Though our system is focusing the limited user layer such as earth environmental researchers, more than 5000 hits per month describe the practical usefulness of it. The distribution of the contents of accessed data shows the demand for various kind of those data.

1. はじめに

昨今の地球環境への関心の高まりとともに, リモートセンシングデータをはじめとするさまざまな地球環境データへの需要が高まっている. しかし, それらのデータ種類・形式は多岐に渡るため, 米国ゴア副大統領も指摘している如く[1], 多くの研究者はその膨大な生データに圧倒されているのが現状であり, 有益なデータの多くは未利用のまま眠っているとさえ言われている.

† 東京大学大学院工学系研究科
Graduate School of Engineering, University of Tokyo
☆ 現在, 日本学術振興会特別研究員
Presently with JSPS Fellow, Institute of Industrial Science, University of Tokyo
†† 東京大学生産技術研究所
Institute of Industrial Science, University of Tokyo

このような背景から, これら地球環境データを一元的に管理し, ネットワークを経由して自由に取得・閲覧可能なデジタルライブラリの研究開発と Web 上での公開が行われつつある[2)3)4)6)7)8)] が, これらのデータの利用者である地球環境工学分野の研究者にとっては, 未だ十分実用的に利用可能なシステムとは言えないのが現状である.

その原因としては次の 2 点が考えられる.1 点は実用的に利用可能なデジタルライブラリへのデータローディングツールの欠如に起因する利用可能データ数の不足にあり, 他の 1 点は旧来の ftp や gopher サービスで用いられていた原始的な手法が中心である検索インターフェースの操作性にある.

そこで本研究では, 実用的な地球環境デジタルライブラリ構築に必要とされる以下の 2 点に於いて有効な手法

の提案と検討を行った.

- 多様なフォーマットを有する土壌・地表面気候等地球環境データに対するデジタルライブラリへのデータローディングツールの開発.
- 容易な操作で必要とするデータの検索が可能な地球環境情報の為のユーザインターフェースの開発

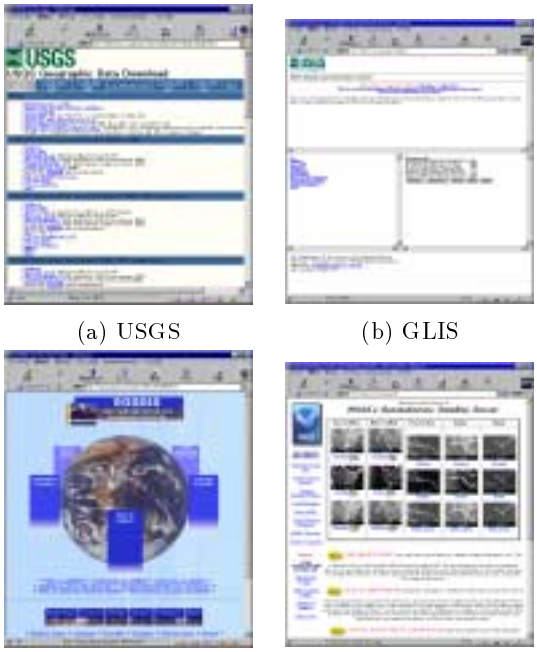
以下,2章で地球環境デジタルライブラリの概要と現在公開されているシステムの特徴, 現状の問題点の明確化を行う. 3章では本研究で提案するデジタルライブラリへのデータローディング手法の特徴と約 1100 種類の土壌・地表面気候データに対する実験結果を示し, 4章では検索インターフェースの特徴と利用について述べる. 5章で現在までの利用実績とシステム構成を示し,6章で結論と今後の課題を述べる.

2. 地球環境デジタルライブラリ

2.1 地球環境データの特徴

本研究で対象とした地球環境データは, 次のような特徴を持つ.

- データフォーマットの多様性
  リモートセンシングデータを始めとする地球環境データは, データ取得手法の多様化に伴い現在非常に多方面で収集・蓄積が行われている. しかし, そのデータフォーマットの統一的な基準が存在しないため, 各組織ごとに独自の様式で管理を行っているのが現状である. そのため, デジタルライブラリの構築に際しては, その各データのローディング作業に大幅な労力が必要とされている.
- 時間属性および空間属性の重要性
  多くの場合地球環境データは属性としてその内容属性に加え, データが対象とする位置属性と, 取得あるいは算出された時間属性を同時に持つ. さらに時間属性に関しては, ある地域の月毎平均気温データなど, 同一地域における継続的な時系列データとして取得されている場合が多い. そのため, 検索時には内容・空間・時間の 3 属性に関する条件設定が可能なインターフェースの設計が必要とされる. また, 視覚化時には, 時間に関する動的変化の把握などが容易な手法の検討が重要である.
- 異なる観測データ間の相関性
  地球環境データは地球上のある地域における諸現象を数値化したものであり, これらの現象は, 例えば同一地域の降水量と気圧の関係のように非常に高い相関性を持つものも多い. また, 同一内容を異なった手法によって算出しデータを作成する場合も多く, データ間の差異は重要な要素である. そのため, データの比較を支援する視覚化手法の検討が必要である.



(a) USGS (b) GLIS

(c) EOSDIS (d) NOAA

図 1 米国の地球環境デジタルライブラリ関連ページ

上述のように地球環境データは, 一般のデジタルライブラリが対象とする文書データや数値データと異なった特性を有し, これらを考慮した地球環境情報指向の新たな手法が必要とされる.

2.2 現在公開されている地球環境情報サイト

現在, 地球環境データを収集したデジタルライブラリが多く Web 上で公開されているが, 本節では図 1に示す代表的なサイトの特徴の分析を行い, その概要を表 1に示す. データソースについては, システムを運用している機関あるいはその協力機関が収集したデータを公開しているタイプは「自組織」と表記し, 外部の機関のデータをデータローディングツールなどを用いて取得し公開しているタイプを「他組織」と記述している.

米国 USGS が公開している US GeoData[6] では, あらかじめ用意されたさまざまな解像度のデジタルマップをベクター形式およびラスター形式で公開しており, ユーザには ftp サービスに準じた形式でのアクセスインターフェースが提供されている. しかし, 検索機能は存在せず, データ一覧時にアルファベット順や緯度経度順でのソートのみ可能である.

NASA EROS データセンターが公開する EOSDIS (Earth Observing System Data and Information System)[8] は NOAA,Landsat,SAR など約 10 種の衛

システム名 (運用機関)	データソース	内容検索	空間検索	時間検索	複合検索	備考
US GeoData(USGS)	自組織	△ (ツリーのみ)	○	×	×	ベクター形式も公開
EOSDIS(NASA-EROS)	自組織・他組織	○	○	△ (10 日単位)	×	
GSS(NOAA)	自組織	○	×	○	×	縮小画像のみ
地球環境 DB(国立環境研)	他組織	△ (ツリーのみ)	×	×	×	CD の郵送が中心
衛星データ検索システム (千葉大)	自組織	△ (衛星のみ)	×	○	×	

表 1 各システムの持つ検索機能

星のデータを収集・公開している. このサイトでは, まずデータの種類を選び, 関心のある地域の緯度および経度を指定することにより, その領域を含む 10 日単位のデータの一覧が表示される. ユーザはその中から関心のあるファイルのダウンロードが可能であるが, データ種類, 位置, 時間という固定された手順での指定のみが可能であり, その他の手法は提供されていない.

NOAA の運営する Geostationary Satellite Server[18] では,NOAA, GOES,METEOSAT,GMS の最新データをクイックルック画像一覧から選択・表示が可能である. しかし, 過去のデータに関しては, 各データごとに観測時刻による検索を行うことでダウンロードが可能となるのみであり, 視覚化支援などは行われていない.

その他, 日本でも高知大学理学部情報科学科[19], 千葉大学環境リモートセンシングセンター[20], 国立環境研[21] などが同種のサービスを行っているが, 多くの場合データ内容や時刻など, 単一の条件か固定された順序での検索しか提供されていないのが現状である.

2.3 問題検討

前節でいくつかの代表的なデジタルライブラリの特徴を紹介したが, いずれも地球環境工学分野の研究においては未だ実用的に利用されていないのが現状である.

その直接的な要因として,

- 利用可能なデータ種類数の不足
- 検索時の容易な操作性の欠如

の 2 点が挙げられる.

第 1 の要因については, 多様なデータをデジタルライブラリに導入するためのローディングツールの欠如に起因すると考えられる. これら地球環境データは,GCM(Global Climate Model) などに代表される将来予測を行うアプリケーションの初期値として用いられる場合が多いが, 例えば現在水文分野で最も信頼性が高いため幅広く用いられている SiB2(Simple Biosphere Model 2) の場合, 植生関連や土壌関連など約 80 種類の地球環境データが必要とされる. しかし, 各機関で取得されたデータは固有の異なったフォーマットを持つ事が多く, 画一的なデータローディングツールを用いたデータ導入は困難である. そのため, 各フォーマットに対応したツールを人手で準備しているのが現状であり, データ種類の拡充には膨大な労力が必要となる.

第 2 の要因については, 現行のシステムでは従来の ftp や gopher サービス上で用いられていたディレクトリ形式によるインターフェースや, 時間や位置などいずれか 1 条件を固定した順序での検索条件指定手法しか提供されていないことによる. 研究レベルにおいて利用する場合には, 複合条件の柔軟な組合せによる検索など, より高度な設定を容易な操作性で利用可能であることが求められるため, 現状システムは満足のゆくものとは言い難い. また, 現状のシステムの多くはその検索結果表示を各データの識別子 ID の一覧によって表示しているに過ぎず, 検索結果の概要把握や妥当性の判断に用いる必ずしも利用が容易とは言えない.

3. 多様な形式の地球環境データを対象としたデータローディングツールの開発

本研究では, 多様な地球環境データを効率的に導入する手法として, 地球環境データの特徴を用い対象データのフォーマットを自動認識することにより, 可能な限り自動的にデータ導入を行うツールの開発を試みた. 本章では, 地球環境データを対象としたデジタルライブラリへのデータローディングツールを提案し, 土壌および地表面気候に関する実際のデータに対して適用した結果について報告する.

3.1 現状と設計指針

前章で述べた通り, 地球環境データのデジタルライブラリへの導入に関しては, 現時点では有効なデータローディングツールが存在せず人手による処理が行われている. 独 ART + COM[5] や米 USGS[6] などでは, 同一フォーマットに従ったデータごとに分類を行った後, 手作業でフィルタを作成し適用することにより導入を行っている. そのため, 定期的に各サイトの調査を行っても新たに拡充されたデータの発見は少なく,GCM などでの利用に対応できる豊富なデータ種類を持つシステムが存在しないのが現状である.

そこで本研究では, 地球環境データを手作業に依存せず可能な限り自動でデータ導入を行うことにより, 多様なデータを持つ地球環境デジタルライブラリの構築が今

後重要であるとの認識から, 次に示す地球環境データのローディングツールの開発を行うこととした.

3.2 デジタルライブラリへのデータローディングツールの試作

本節では, 本研究で構築するデジタルライブラリを構築する上で, 地球環境データから抽出する必要があるデータの属性とその用途を述べ, その抽出過程において利用する地球環境データの特徴と具体的なローディング処理手法を示す.

3.2.1 デジタルライブラリ登録属性とその用途

地球環境デジタルライブラリにデータをローディングし, ユーザが利用可能とするには, 次のような情報が必要とされる.

- 形式の変換
  - ファイル形式
  - データ構造
- 属性の抽出
  - 日時情報
  - 内容情報
  - 空間情報
  - データの値に関する情報

ファイル形式に関する情報は, 取得されたファイルをデータとして利用可能とするための圧縮形式や数値形式, 数値型情報などであり, データ構造は各ファイルに含まれているデータを抽出するために必要とされる情報でヘッダ・データ部の構造に関するものなどがある.

データ属性に関しては, 地球環境デジタルライブラリ上でユーザが検索する際に必要となる情報として, データの取得された日時に関する情報, データの内容を示す情報, データの示す位置や範囲など空間に関する情報などがある. また, 位置属性に関しては, 用途に応じ使い分けられた地図投影法に関する情報も必要とされる.

以上の処理で抽出される属性について,RDBMS に格納される属性, 具体例, 主要な用途を表 2に示す.

なお, 観測データは原データならびに等緯度経度座標系に変換したデータが登録される.

3.2.2 認識時に利用するデータの特徴

本研究で対象とする土壌・地表面気候に関する地球環境データに関し, 認識時に利用した特徴は以下の通りである.

- ファイル形式に関する特徴
  地球環境データの多くは時系列のデータセットとして構成されており, 各データは個別にファイルとなっている場合と, すべてが結合されて 1 ファイルになっている場合が存在する. 各データは通常観測データを主構成要素とするが, 更にそれに加えてヘッダ部・フッタ部に示した付随情報が記録されることがある. 観測データ部は数値で構成され, ヘッダ部・フッタ部は文字等が多く含まれる.
- ヘッダ・フッタ部, 観測データ部の特徴
  地球環境データに含まれるデータ部は, ヘッダ・フッタ部に比較してその表記法が非常に単調な数値列である特徴を持つ. ヘッダ・フッタ部に於いては, 空白の出現間隔が一定ではなく, 文字なども多く含まれているが, データ部では数値や小数点のみで構成されている場合がほとんどである.
- 観測データ値の特徴
  - 隣接点間の値変化
    一般に地球環境データは, 隣接地点の値との差が小さい傾向がある. これは, 多くのデータは地球上で連続的な自然現象を示しているため, 陸域を示す値と海域を示す値間, すなわち海岸線を挟む場合を除き極端な変化を示さない.
  - 時間表記
    地球環境データに含まれる時間属性を示す表記は, 各データの取得が 3 時間毎, 10 日毎などの取得間隔が記録されることが多い

属性	例 1	例 2	用途例
データ取得機関	gswp	CTR	内容検索
データ分類名	NCEP		内容検索
データ名称	Fld	WOA	内容検索
始点緯度	-90	-35	空間検索, 視覚化処理
始点経度	-180	-10	空間検索, 視覚化処理
終点緯度	90	30	空間検索, 視覚化処理
終点経度	180	60	空間検索, 視覚化処理
画素数縦	360	100	空間検索, 視覚化処理
画素数横	180	100	空間検索, 視覚化処理
日時	870101	9401	時間検索, 動画作成
期間	1 day	1month	時間検索, 動画作成
欠損表記	-9999	-32767	視覚化処理
投影法	等緯度経度	正距方位	データ導入処理, ダウンロード
数値型	整数	小数	データ導入処理
最大値	132	-23.4	データ導入処理, 視覚化処理
最小値	51	-48.1	データ導入処理, 視覚化処理
既登録キーワード	mmHg		内容検索
その他	Global Soil Wetness Project	Sea Temp Depth 1000	内容検索
等緯度経度観測データ格納ディレクトリ	/home/data/gswp/ncep/fld	/home/data/CTR/woa	ダウンロード 検索・視覚化処理
原データ格納ディレクトリ	/home/data/gswp/ncep/org/fld	/home/data/CTR/org/woa	ダウンロード

表 2 各データから抽出される属性とその用途例

– 欠損値表記
0000 や 9999 など自然現象を示したデータ中にはほとんど存在しない値で表記されている.
– 画素数
地球環境データを構成する画素数に関しては,1 度あたりの画素数を基準にした表記法 (5pixel 度 など) を用いる場合が多いため, 全球を示す経度値 (360) や緯度値 (180) の整数倍であることが一般的である.

以上のような特徴を用い, 次節に示すデータローディング処理を行った.

3.2.3 認識の流れ

本手法では, 土壌・地表面気候に関する地球環境データの特徴を利用し, 各処理対象データに対し次の順序で処理を行うことにより認識を試みている.

(1) 解凍処理, アスキー変換
(2) ファイルの分割, 属性の分離
(3) 数値型認識
(4) ヘッダ部等データ部以外からの日時情報等の属性抽出
(5) データ部の解析による欠損データ表記法および縦横画素数の決定
(6) 投影法認識, 変換
(7) 位置認識
(8) 検出失敗ファイルのチェック
(9) その他メタ情報の参照

このうち,(1) 〜 (5) まではデータベースへの導入時に処理の成功が必須の項目であり, 失敗した場合にはその時点で処理を中断し自動導入を放棄する. (6)(7) は認識が失敗した場合にもその時点での認識結果に基づいてデジタルライブラリに登録を行う. (8)(9) は自動処理が困難であり, ユーザへの問合せによる手動の処理である.

各処理の具体的な手法は次に述べる.

3.2.4 処 理 手 法

(1) 解凍処理, アスキー変換
本処理では, 最初にデータ圧縮の有無および圧縮方法の推定を行う. 拡張子からの推定, ヘッダ部に記述されたキーワードからの類推, データ形式の認識などを行い, 主に UNIX 系で使われる gzip,compress と Windows 系で用いられる lha,zip 形式の判定と解凍処理を行う. また, 解凍処理を行ったデータに対し, バイナリフォーマットの場合はアスキー変換を行うため, バイトオーダーが Big Endian か Little Endian かの判別を行う. 本手法では,Big Endian である場合と Little Endian である場合のそれぞれの変換ツールを用意し, 該当データに適用する. その結果のアスキーデータに対し, 複数のサンプル地点の値を抽出し, 隣接地点の値との比較を行う. 地球環境データの場合, 陸域と水域の境界など稀な場合を除き, 隣接点の差異は小さい傾向があるため, 両結果で差異が小さい処理手法のバイトオーダーと判定し, アスキー形式への変換結果とする.

(2) ファイルの分割, ヘッダの分離
本処理では, 複数のデータが結合されたファイルの分割を行う. 一般の地球環境データのデータ部の文字列はヘッダ部に比較して単調な表現形式である点を利用し, 規則的な数値列が文字列を含む (ヘッダと類推される) 部分を挟んで 3 度以上出現した時点でファイルの分割の必要性を判定する. ファイルの先頭から単調表現部分の先頭直前までをヘッダ部とし, そのサイズをヘッダサイズとする. また,2 度目に出現するヘッダ／フッタ境界部から 3 度目の境界部までの領域をフッタ部とヘッタ部の結合領域とし, 前述のヘッダサイズを用いてフッタ部の確定を行う.

(3) データの数値型判定
整数および小数, 正負符号の有無を各データの記述形式から判定を行う. 本処理では全データの走査を行うため, 同時に最大値および最小値の検出も行う.

(4) ヘッダ部等データ部以外からの日時情報等の属性抽出
本処理では, 表示形式や最大値など, データ部で認識された属性以外で, 主にヘッダ部から抽出可能な属性の認識を行う. 本手法では, 地球環境データの次の 5 つの属性情報抽出処理を試みている.

- 年月日時情報
予め作成した典型的な表記パターンテーブルとの比較により, 数値列中から年, 月, 日, 時刻の推定をで行う. 年や月を示す数値である可能性を有する数値範囲および表記方法との対照を行い, さらに時刻における「:」など特有の表記方法を検出して認識を行う. また, 年月日時情報が記述されていると推定される箇所と同一の箇所を, 同時に取得した他ファイルから抽出し, 各数値の変化量が一定の範囲内に含まれると同時に各データに周期性が類推されれば年月日時情報と決定する.
- 期間
前述のテーブルとの比較により, ヘッダ部から 24(時間),7(日 / 週),3,6,12 など期間を表

現していると推定される数値を抽出する．また, 前述の年月日時情報のデータ間による差分情報も利用し判定する．

- 欠損データ表記法
  本処理では, データ中に存在する欠損データ表記法の推定を行う．本手法では,「9999」「0000」など地球環境データにおける典型的な欠損データ表記法を登録した参照テーブルと, 本手法を適用して処理を行った際に, 欠損データとして認識された値を登録する履歴テーブルの 2 つを保持し, ヘッダ中から両テーブルに含まれる値の検索を行い, 該当するデータを欠損データ表記法の候補とする．この決定は後述のデータ部解析時に行う．
- 縦横画素数推定
  一般に地球環境データは緯度方向と経度方向について格子状に配列された構造を持っており, 本研究ではそれぞれ緯度方向のデータ数を横画素数, 経度方向のデータ数を縦画素数と呼ぶ．本処理では, ヘッダ部中の検索を行い, 地球環境データにおいて典型的な画素数表記である 180(度) の倍数および約数, あるいは 2:1 の比を持つ数値列を検索し, 縦横画素数の候補とする．この決定も後述のデータ部解析で行う．
- その他属性
  数値単位や Copyright 表示など, 典型的な表記形態を表記パターンテーブルとの比較によって抽出する．

(5) データ部解析による欠損データ表記法および縦横画素数の決定
本処理では, 前処理で認識を行った欠損データ表記法および縦横画素数の推定候補中から, データ部分の解析によって各値を決定する．

- 縦横画素数の決定
  前処理で候補が抽出されている場合は, データ部の項目数との比較による決定を行う．また, 改行コードが存在する場合は, 改行コード間の項目数を横方向画素数と類推し前述の候補と比較を行う．両者とも存在しない場合, この段階で自動処理を中止する．
- 欠損データ表記法の決定
  最初に, 前処理で推定された候補値から, データ部において最も出現頻度が高い値を抽出する．この値が参照・履歴テーブル内に含まれる値の場合は欠損データ表記法として決定する．含まれない場合においても, データ部の値の分布から著しく逸脱し高頻度な値が 1 つだけ存在する場合には欠損データ表記法と決定し, 新たに履歴テーブルへの登録を行う．高頻度な値が 2 つ以上出現する場合は本処理を停止し, この段階で自動処理を中止する．

(6) 投影法認識, 変換
本処理では, 地球環境データの投影法の類推と等緯度経度座標系への変換を行う．対象データの全点に値が存在しない場合には, 特徴的な輪郭形状を持つ正距方位図法やボンヌ図法, グード図法の各図法から等緯度経度座標系への変換処理を行い, 処理データの輪郭形状を等緯度経度座標系の輪郭と比較を行い推定する．全点にデータが入力されている場合, 輪郭形状での類推は不可能のため等緯度経度座標系か一部領域のデータと推定する．この決定は (7) 位置認識処理で行う．

(7) 位置認識
本処理では前処理で行った等緯度経度座標系への変換後, 地球環境データの地球上における対応位置の認識を行う．最初に等緯度経度座標系に投影した全球の海岸線データと比較を行う．45 度単位で南北・東西方向に移動を行い, 最も海岸線と一致する位置で決定し, 緯度経度 0 度を中心座標とする系に移動する．一致しない場合, 部分領域の可能性を考慮し,1 度単位での移動を行い対応地点を検索する．この処理でも対応地点を発見できない場合は位置認識を中止する．

以上の処理を行って抽出された地球環境データの属性情報を, 等緯度経度座標系に変換したデータとともにデジタルライブラリに登録する．

3.3 実験結果

前節で述べたデジタルライブラリへのデータローディング手法を用い, 本研究では降水量, 雲量, 湿度など気候に関した観測による約 600 種類のデータと, 土壌に関するモデルによって生成された約 500 種類のデータを対象とし, データローディングのためのフォーマット認識実験とデータに含まれる属性情報抽出の実験を行った．

これらのデータは, 表 3に示す約 30000 ファイル, 圧縮状態で約 20GB に達する膨大なものであり, そのデータフォーマットおよび空間時間解像度が非常に多岐に渡るデータセットである．

3.3.1 フォーマット認識実験

本研究で開発したデータローディングツールを用い, 地球環境デジタルライブラリへローディングを行うためのフォーマットの認識実験を行った．

Global Data Sets for Land-Atmosphere Models	植生, 土壌, 雪氷, 雲, 放射量, 海水面温度, 湿度など多様な分野の地表面気候に関するデータを元に生成された約 600 種類 20000 ファイルに及ぶデータセット. 時間解像度は 3 時間〜 1 カ月, 各 2 年分.
Global Soil Wetness Project Data	上述 Global Data Sets for Land-Atmosphere Models を用いて, 本プロジェクトに参加している 10 の研究機関が土壌水分などに関し各自のポリシーの基づいたモデルに従い算出したデータ集. 時間解像度は 3 時間〜 1 カ月, 各 2 年分. 約 500 種類,10000 ファイル.

表 3 実験対象データ

認識処理内容	対象数	認識数	認識率
解凍処理, アスキー変換	28921	28813	99.9
ファイルの分割, ヘッダの分離	28813	25319	87.9
数値型認識	25319	24991	98.7
縦横画素数の決定	24991	24051	96.2
投影法認識	24051	20372	84.7
位置認識	20372	20160	98.8

表 4 フォーマット認識率

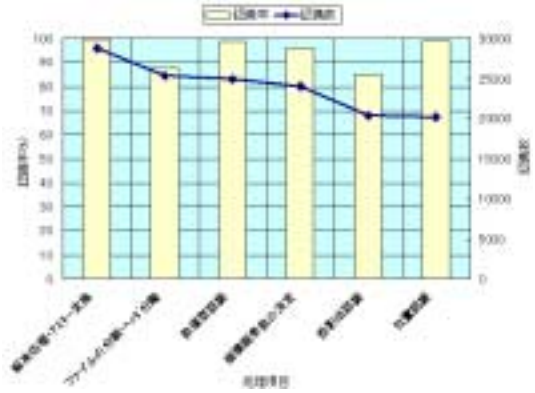


図 2 認識率の推移

データベースへのローディングには, 前節で示した認識処理のうち (1) 〜 (3) および (4)(5) から導出される縦横画素数が必要であり, 認識が必須ではない (6)(7) の処理も含めたこれら 7 つの処理の認識率を表 4に示す. また, この 7 段階の処理において 1 段階前の認識が成功したものに対してのみ次の処理を行った結果について, 認識率の推移を図 2に示す.

3.3.2 属性抽出実験

本ローディングツールを用い, 前節で示した認識の流れの (4)(5) の段階で行われるヘッダ部・ファイル名・データ部からの属性抽出実験を行った. 認識処理を行う対象データならびにその属性と認識が成功した認識数を表 5に示す.

ヘッダ部からの属性抽出実験は, フォーマット認識実験のファイル分離・ヘッダの分割処理が完了した 25319 データのうち, ヘッダ部が抽出できた 3815 データを対象として行った. ファイル名からの属性抽出実験は, 前実験の解凍処理・アスキー変換が完了した 28813 データを対象とし, データ部の解析による属性実験は数値型認識の処理が完了した 24991 データを対象としている.

属性	認識数	認識処理を行う対象データ (対象データ数)
年月日時	731	ヘッダ部から (対象データ:3815)
期間	3025	
縦横画素数候補	3211	
欠損表記法	201	
年月日時	21031	ファイル名から (対象データ:28813)
圧縮形式	1339	
欠損表記	3302	データ構造から (対象データ:24991)

表 5 属性抽出数

3.3.3 検 討

本研究で開発したデータローディングツールを適用することにより, デジタルライブラリ登録のために必須である縦横画素数認識までは約 80%, 位置認識まで含めた認識処理は約 70% の認識率で自動認識が実現された. フォーマット認識実験に関しては, 投影法認識の認識率が表 4で示すように最も低い値になっている. 現状では正距方位図法, ボンヌ図法, グード図法をサポートしているが, さらに機能拡張を行うことにより改善されると予想され, 今後実装を進めてゆく予定である. 属性抽出実験における年月日時の抽出はファイル名からが基本であるが, 一部はヘッダ部からの認識も可能であることが分かる. 今後はデータセットに付随するドキュメントファイルの解析等も行うことにより, さらに多くの属性情報を抽出する手法を検討する予定である.

4. 検索インターフェースの検討

本章では, 現状のシステムにおける 2 番目の問題点である検索インターフェースについて, 本研究で提案する内容・空間・時間に関する 3 方向からの検索手法の特徴と実装結果を示す.

4.1 現状と設計指針

2章で述べたように, 現在公開されている地球環境デジタルライブラリを調査すると, 空間や時間など単一の条件での検索のみ可能である場合が殆どであり, 実研究分野において需要の高いこれらの複合した検索を行う場合の容易な手法が提供されていない.

そこで本研究では, 複数の条件に関して自由に切り替えながらの検索が容易に実行可能であり, その検索結果

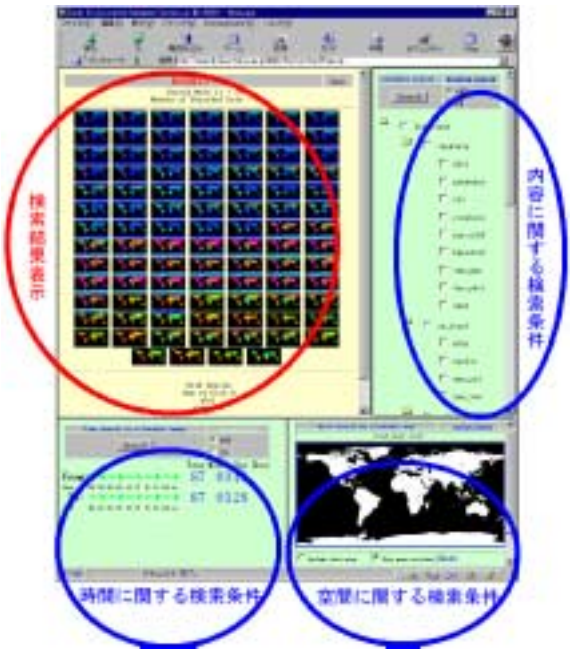


図 3 3 情報に基づいた検索ページ

を確認しながらより多くの検索を連続して実行が可能な検索手法の提案を行う.

その有用性を実証するために, 地球環境工学分野の研究者との詳細な検討の結果, 以下の 3 種の情報に基づいた検索を任意に組み合わせ可能なて行うインターフェースを Web 上に実装した.

- 内容情報に基づいた検索
- 空間情報に基づいた検索
- 時間情報に基づいた検索

具体的な実装手法および利用法を次節に示す.

4.2 内容, 空間, 時間の 3 方向からの検索

本インターフェースの実装に際しては, 内容, 空間, 時間の各情報検索インターフェースを図 3のように配置した. 左上のフレームに検索結果を表示するウィンドウがあり, その周囲に内容, 空間, 時間情報に関する検索ウィンドウが配置されている.

ユーザが各検索ウィンドウで条件指定を行うと同時に, その内容に従った SQL が生成され, 本システムの検索部で検索された結果が結果表示ウィンドウに表示される. 1 つの検索ウィンドウで条件を指定後, 別の検索ウィンドウで指定することにより結果出力ウィンドウ上の検索結果はさらに絞り込まれた検索結果が表示される. この操作は任意の順序で可能であるため, ユーザは各検索条件を自由に行き来しながら必要とするデータの検索を進めることが可能である.

Web ブラウザ上に実装された本手法のインターフェースを用いることにより, 従来の手法に比べ各条件による検索結果を確認しながら更に検索を行うことが可能であるなど, 条件数およびその順序においてより柔軟な検索が実現された. 各情報に基づいた検索インターフェースの特徴は次の通りである.

4.2.1 内容情報に基づいた検索

地球環境データの検索の際に最も一般的な手法は, データの内容情報に基づいた検索である. 各データの取得機関やその分類, データ名についての条件を設定し検索する手法であり, 本システムではツリー構造による検索とキーワードからの検索の 2 手法のインターフェースを実装している.

キーワードによる検索は, そのキーワードが含まれるデータ種類名を検索し, 該当データがその分類情報とともに選択肢として表示される. その中から必要なデータにチェックを入れて検索することにより結果表示ウィンドウに該当データが表示される.

本手法を導入することにより, 例えばキーワード検索を用いて他組織で収集された同種のデータを同時に閲覧して比較するなど, より実用的な利用が可能となった.

また, ツリー構造からの検索ページでは, 格納されているデータが属性情報に基づいて階層的に表示されるため, ユーザは必要に応じた階層での選択を行い検索を実行する.

4.2.2 空間情報に基づいた検索

空間情報による検索ウィンドウでは, 緯度経度による検索とクリッカブルマップによる検索の 2 種類のインターフェースを実装している.

緯度経度による検索は, 検索する領域の左上座標および右下座標の入力を行い, 南北緯, 東西経の区別を選択して検索を行う. あらかじめ検索対象領域が明確に分かっている場合など, 領域を正確に指定したい際に有用である.

クリッカブルマップによる検索は, 表示されている地図上で検索領域の左上座標と右下座標を順にクリックすることにより指定が行われる手法を実装した. 本手法では, 最初のクリックで左上座標が確定されるため, その点より右下に位置する領域が拡大され表示される. 次のクリックでは同様に右下座標が確定されるため, 前回のクリック点とで定義される領域が同様に拡大され表示される. 以降, 繰り返し左上座標と右下座標の指定を交互に行うことにより, より高解像度な地図上での領域指定が可能となる.

4.2.3 時間情報に基づいた検索

時間情報を用いた検索に関しても,2 つの手法を提供している. 数値を入力することにより検索開始日時および

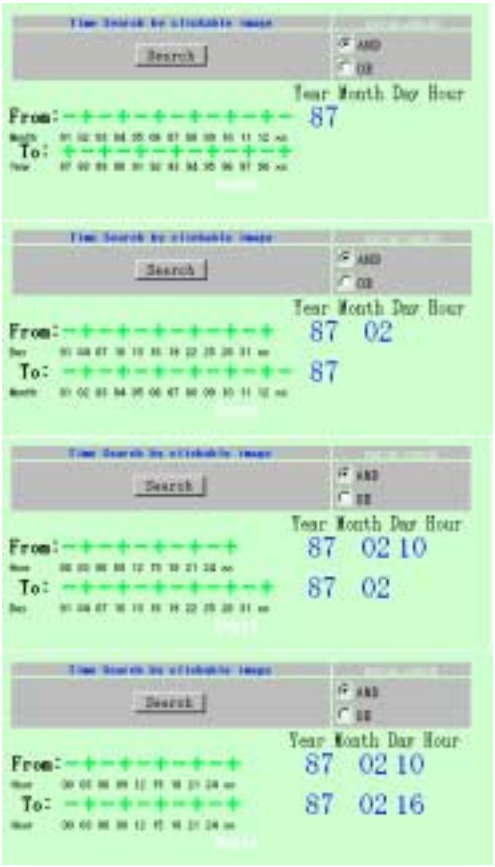


図 4 マウスによる時間指定検索

終了日時を指定する手法と, 図 4に示すように, 表示された時間軸をマウスで選択していくことにより指定する手法である.

マウスによる指定の場合, 最初に開始年に該当するところをクリックすると, その値が右に表示され, 開始時間軸は次の階層である開始月が表示される. 同様にクリックすることにより開始日が表示され, 順に詳細な時間の設定が行える. この際, 終了時刻を示す時間軸は, 初期入力時を除いて開始時間軸の一階層上位の時間軸を示している. すなわち, 開始軸で開始日を指定する際には, 終了軸には月を表示した軸が表示されている. すなわち各入力時には, 開始時刻のより詳細な階層の指定か, 同一階層での終了時刻の指定かを選択する.

両手法において,「＊＊」を選択することによりその階層の値を任意とすることが可能である.

4.3 検索結果の視覚化

4.3.1 検索結果の一覧

各検索ウィンドウで指定された条件に基づいた検索結果表示に関しては, 従来, データの識別子を表示していたが, 本研究では検索結果データそれぞれを画像化したも

(a) 検索結果数＝ 96

(b) 検索結果数＝ 6

図 5 結果出力ウィンドウ

のを結果表示ウィンドウに一覧表示することにより, 利用者が全体を容易に把握出来る様工夫した.

なお, 表示画像は, その分類に属する全データにおける最大値および最小値に基づいて正規化したカラーチャートを用い, 各値を色で示した.

結果出力ウィンドウ上では, 検索該当データ数に応じ,1画面中に全データが表示されるように各データのサイズが動的に変更され表示される (図 5).

地球環境研究者は概要を視覚化して把握することが可能となり, 大きく利便性を改善することができた.

VRML を用いることにより, ユーザは仮想空間中に

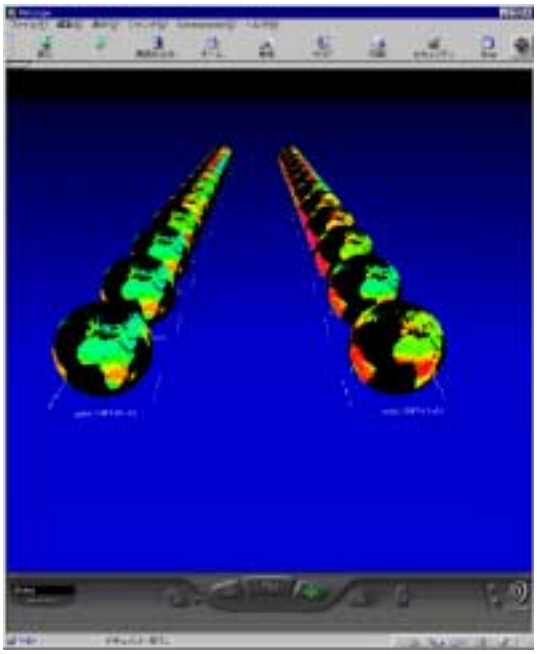


図 6 時系列に整列した配置

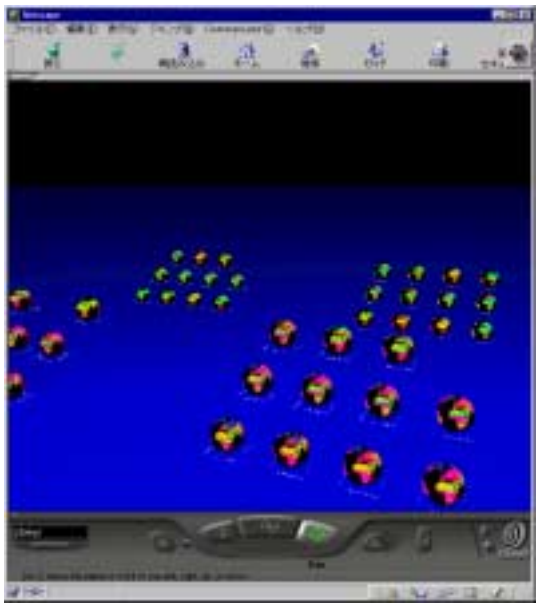


図 7 内容に応じ分散した配置

配置された各検索結果データを示す VRML オブジェクト間をウォークスルーしながら閲覧を行い, 関心のあるデータには自由に接近し任意の角度・距離からの確認が可能である.

また, 容易に時間による変化を確認する手法として同一時刻のデータが隣り合うような配置 (図 6), データの内容ごとに分類して閲覧するインターフェース (図 7) を提供している.

4.3.2 詳細なデータ視覚化

最終的に関心のあるデータが確定した場合, 結果表示画面上の縮小画像か VRML 空間中のデータをクリックすることにより図 8 ような当該データの詳細情報ページが表示される. このページでは以下の利用が可能である.

- オリジナル, 並びに等緯度経度観測データのダウンロード
- 原解像度での 2 次元画像表示
- VRML による 3 次元表示
- 2 次元画像の時系列アニメーション表示
- VRML による時系列アニメーション表示

GCM など外部アプリケーションの入力データとして利用する際には, 本ページ上のオリジナルデータが利用される場合が多い. また,2 次元画像および VRML による時系列アニメーション表示は, 同種の時系列データ閲覧手法として効率的であり, 諸現象の動的変化の把握にも有効である.

なお,VRML ならびにアニメーションに用いられるデータは, デジタルライブラリ中には格納されず, 等緯度経度観測データより動的に生成される.

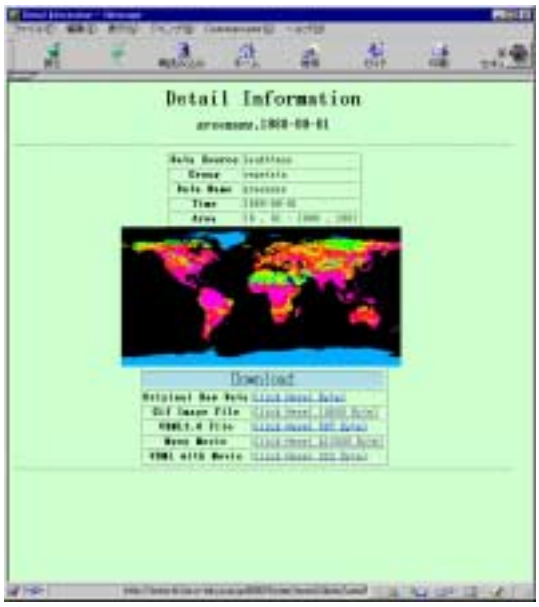


図 8 詳細表示ウィンドウ

5. 利用実績とシステム構成

5.1 利 用 実 績

本研究において実装した地球環境デジタルライブラリシステムは, http://www.tkl.iis.u-tokyo.ac.jp:8080/DV/ において運用を行っている. 1998 年 12 月の

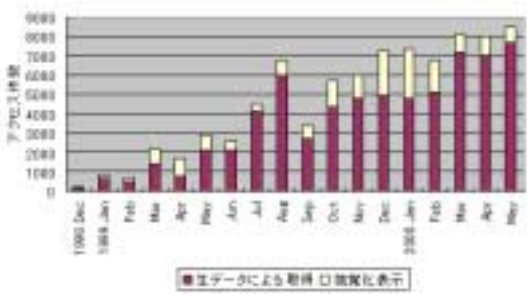


図 9 月別アクセス数

全アクセスに占める割合 (%)	ドメイン	国名
50.4	jp	Japan
12.1	(unresolved)	不明
8.7	edu	US Educational
6.9	uk	United Kingdom
5.1	kr	Korea (South)
3.5	com	US Commercial
3.2	sg	Singapore
2.8	at	Austria
2.6	th	Thailand
1.7	net	US Network
0.9	org	US NP-Organization
0.7	fr	France
0.6	it	Italy

表 6 トップドメイン別アクセス数割合

データ処理サーバ	Sun Enterprise 6500 (UltraSPARC-300MHz × 6,2GB Memory)
Web サーバ	Sun StorEdge A5000(250GB)
大規模記憶装置	OS:Solaris 2.6
ソフトウェア	httpd:Apache httpd-1.3.0
	DBMS: IBM DB2 V6.1

表 7 用いたハードウェア・ソフトウェア

公開以来の利用実績を図 9に示す. 専門的なデータにもかかわらず利用者数は増加傾向にあることが判る.

また, 表 6に示すように海外からのアクセスが約 3 割を占めるなど, 国内に限らず幅広く利用されていることが分かる.

5.2 システム構成

本デジタルライブラリシステムは, 大きくデータローディング部, データベース部, インターフェース部の 3 部から構成されている.

データベース部には IBM 社の RDBMS である DB2 V6.1 を用いており, システムを構成するハードウェア・ソフトウェアは表 7の通りである.

6. おわりに

本研究では, 近年多方面における需要が増加している地球環境データを対象としたデジタルライブラリシステムの構築において, 従来システムの問題点の検討を行い, データ導入時のデータローディングツールの開発とデータ検索インターフェースについて新たな手法の提案を行った.

現在, データウエアハウスが広く利用されるにつれて, 種々のデータローディングツールの商用化が進みつつある. これらは主として, ビジネスデータを対象としており各種 DBMS, ならびに ERP 等からのデータ抽出を容易化している.

これに対し, 本研究では対象を土壌, 地表面気候分野の地球環境データを対象とし, そのローディングの自動化, 高効率化を目指しツールの開発を試みた. 対象を限定することにより実用レベルの認識率を得ている. 適応領域を拡大することにより, 種々の課題が生じることが予想され, 今後の課題としたい.

さらに, 内容・空間・時間属性の 3 方向からの検索が可能であり,VRML を空間中のウォークスルーを用いたデータ閲覧手法を導入した検索インターフェースの Web 上への実装を行い, 容易な操作で柔軟な検索を実現した.

本インターフェースは, 従来のシステムでは条件指定の柔軟性に問題があったとの地球環境工学分野の研究者の意見に基づき, より実研究における利用形態に即した操作性を実現している. 各専門分野に研究者においては, データの概要把握にはテキストで示される属性情報よりもむしろ画像による一覧が有効であるという意見に基づき, 縮小画像のみの検索結果一覧手法を試みている. 本システムを基礎として, 地球環境アプリケーション連携型システム SiB2 on Web[22] を構築し運用を行ったところ, 多くの専門家によって実利用され多くの知見が見出されつつあり[23], その観点からも有効性を明らかにした.

これらの手法を実装した地球環境デジタルライブラリシステムを一般に公開し運用を行った結果, 地球環境工学という限られたユーザ層を対象としているにもかかわらず, 毎月 8000 件以上のアクセスを記録し, 本システムの実用性の高さを示した. また, アクセスされたデータ種類も多岐に渡り, 広範なデータへの需要が高いことを実証した.

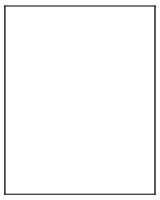
謝 辞

本研究を行うにあたり, 繰り返し御討論頂きユーザの視点から数多くの貴重な御意見を頂いた未来開拓学術研究グループ「水・物質バランスの時空間変化に着目した人間活動の環境影響評価とその軽減方策に関するシステム研究」 (代表：虫明功臣) の皆様ならびに東京大学生産技術研究所第 5 部虫明・沖研究室の新井崇之氏, 金 元植氏に深く感謝致します.

参 考 文 献

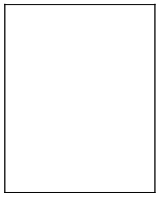
1) Al Gore, “The Digital Earth: Understanding our planet in the 21th Century”, http://www.digitalearth.gov/speech.html.
2) USGS, “Digital Earth Explorer”, http://dss1.er.usgs.gov
3) Cornell Univ. INSTOC, “Cornell’s Digital Earth”, http://atlas.geo.cornell.edu/
4) Microsoft, “Microsoft TerraServer”, http://terraserver.microsoft.com/
5) ART+COM, “Terra Vision”, http://www.artcom.de/project/t_vision/
6) USGS,“US GeoData HomePage”,http://edcwww.cr.usgs.gov/doc/edchome/ndcdb/ndcdb.html
7) SGS,“USGS EDC Global Land Information System (GLIS)”, http://edcwww.cr.usgs.gov/glis/glis.html
8) EROS Data Center Distributed Active Archive Center (EDC DAAC) , “NASA’s Earth Observing System Data and Information System(EOSDIS)”, http://edcwww.cr.usgs.gov/landdaac/
9) Jeanne Behnke,Alla Lake:“EOSDIS: Archive and Distribution Systems in the Year 2000”, Proceeding of 8th NASA Goddard Conference, pp.313-324, Mar.2000.
10) M.Takagi, “Data Reception, Processing, Distribution and Archives Activities at the Institute of Industrial Science, University of Tokyo”, Third AVHRR Users Meeting, Oxford, Dec. 1987.
11) M.Takagi, “NOAA Satellite Data Reception and Processing at the Institute of Industrial Science, University of Tokyo”, Proc. of 2nd Korea-Japan Symp. “Environmental Monitoring from space”, Dec. 1993.
12) Kitsuregawa Lab., “Earth Environmental Data Visualization System”, http://www.tkl.iis.u-tokyo.ac.jp:8080/DV/.
13) 生駒 栄司, 沖 大幹, 喜連川 優, “地球環境データ視覚化システムの構築”, 電子情報通信学会第 11 回データ工学ワークショップ (DEWS2000), Mar. 2000.
14) 生駒 栄司, 喜連川 優,“デジタルアース可視化システムの試作”, 情報処理学会第 59 回秋期全国大会予稿集, Vol.3, pp.205-206, Sep. 1999.
15) Eiji Ikoma, Taikan Oki, Masaru Kitsuregawa, “Development of an Earth Environmental Database System which Interacts with Application Software”, Proc. of 1999 International Symposium on Database Applications in Non-Traditional Environments (DANTE99), pp.252-255, Nov. 1999.
16) 生駒 栄司, 喜連川 優,“陸面植生シミュレータと連携した地球環境データベース可視化システムの開発”, 情報処理学会第 120 回データベースシステム研究会研究報告 pp.153-160, Vol.2000, No.10, Jan. 2000.
17) Kitsuregawa Lab., “SiB2 on Web Homepage”, http://www.tkl.iis.u-tokyo.ac.jp:8080/DV/sib2/.
18) NOAA National Environmental Satellite, Data, and Information Service,“NOAA’s Geostationary Satellite Server ”, http://goeshp.wwb.noaa.gov/
19) 高知大学理学部情報科学科,“Kochi Univ. Weather Home”, http://weather.is.kochi-u.ac.jp/
20) 千葉大学環境リモートセンシングセンター,“Satellite Image Archive for Chiba University”, http://ceres7tx.cr.chiba-u.ac.jp:8080/ja/ceres.htm
21) 国立環境研究所 地球環境センター,“地球環境データベース”, http://www-cger.nies.go.jp/index-j.html
22) 生駒 栄司, 新井 崇之, 金 元植, 沖 大幹,“陸面植生モデルワークベンチの開発と熱帯水田観測データの適用”, 水文・水資源学会誌, 第 13 巻第 4 号, pp.291-303, Jul. 2000.
23) 新井 崇之, 金 元植, 沖 大幹, 虫明 功臣,“ 熱帯水田への SiB2 の適用と水田スキームの導入”, 水工学論文集, Vol.44, pp.175-180, Apr. 2000.

(平成12年6月20日受付)
(平成12年9月21日採録)



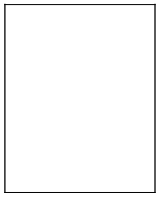
生駒　栄司

昭和 47 年生．平成 7 年東京大学工学部電子情報工学科卒業．平成 12 年同大学院工学系研究科電子情報工学専攻博士課程修了．博士 (工学)．現在, 日本学術振興会特別研究員．地球環境データを対象としたデジタルライブラリに関する研究に従事．



沖 大幹

昭和 39 年生．平成元年東京大学大学院工学系研究科土木工学専攻修士課程修了同年東京大学生産技術研究所助手、現在同助教授。平成 7 年から 9 年にかけて米国 NASA/GSFC 客員科学者。博士 (工学)。地球規模の水循環や水資源、陸面過程のリモートセンシング等の研究に従事 AGU, AMS, IAHS、土木学会、日本水文科学会、水文・水資源学会、日本気象学会各会員．



喜連川 優（正会員）

昭和 30 年生．昭和 53 年東京大学工学部電子工学科卒業．昭和 58 年東京大学大学院工学系研究科電子情報工学専攻博士課程修了．工学博士．同年同大生産技術研究所第 3 部講師．現在同教授．データベース工学の研究に従事．IEEE,ACM 各会員．