大規模観測データの効率利用化に向けた試み

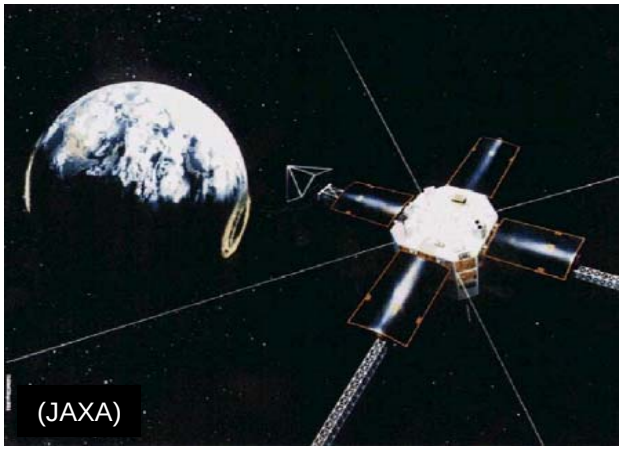
笠原　禎也, 高田　良宏(金沢大学)

(2008.1.10)

金沢大におけるデータ蓄積・利用の現状

あけぼの（1989.2 ～ ）

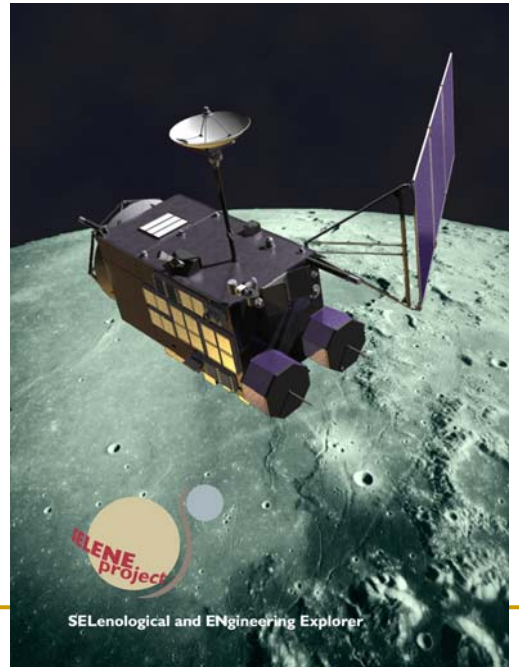
- デジタルデータ: 約2Tbyte
  （うち約1/3が電波計測データ）
- アナログデータ:
  DATテープ約20,000本
  （デジタル化後のデータ総量は
  約20Tbyte）



かぐや（2007.12 ～ ）

- データ伝送量

  全量：10GB/day
  （うち LRS/WFC：1GB/day）



(JAXA)

当研究グループの取組み

- あけぼの/かぐやの自然波動観測データベースの構築
  - Tbyte オーダーの観測データを体系的にデータベース化
  - 観測条件からの最適な観測データの検索・抽出
  - 研究者へのデータ閲覧・配信機構の整備

- TByteオーダーの自然科学DB構築とデータ検索・配信技術の検討

- 大容量の自然科学データに対し汎用的に利用可能なシステムの構築

我々に必要なものは何か？

- テラバイトオーダーのデータの蓄積
  - 大量データのDB化と公開システムの整備
  - 研究者に負担がかからないデータ生成・蓄積機構の構築
  - 誰にでもわかりやすいデータフォーマットの定義

衛星観測データ汎用フォーマット設計上の留意点

- 欠測が多い
- データ容量が大きい

  例：あけぼのWBAでは、生データは約700MB/hour

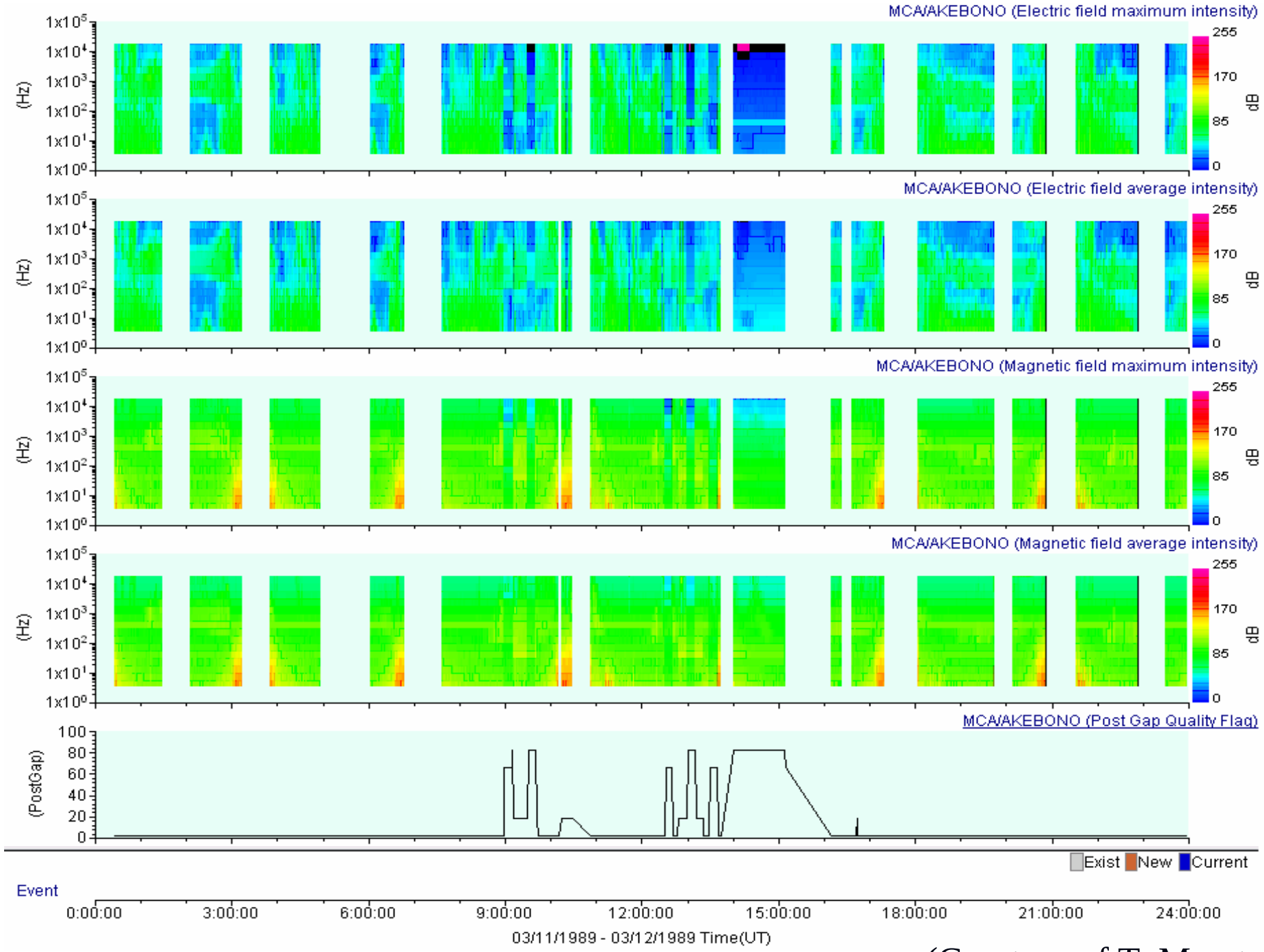
  - データ欠損があってもファイル容量が大きくならない
  - ランダムアクセスが可能

    cf. 従来形式（あけぼのSDB形式）：1可視軌道1ファイル
- 観測モードが複数存在

  時間分解能や測定成分 etc. が観測モードで変化

  - 観測モードに依存するデータフォーマットの違いを（データ処理に影響が出ない限り）なるべく吸収
- 自己記述型汎用フォーマット
  - メタデータをあわせて記述

CDF形式の採用



(Courtesy of T. Murata)

我々に必要なものは何か？

- テラバイトオーダーのデータの蓄積
  - 大量データのDB化と公開システムの整備
  - 研究者に負担がかからないデータ生成・蓄積機構の構築
  - 誰にでもわかりやすいデータフォーマットの定義
  - 適切なデータ公開ポリシーに基づくデータ閲覧機構

データ閲覧システムの構築
（1）開発条件

- 実験・計測データを実際に蓄積・管理する研究室レベルでのデータベース構築を想定
  - データベース・サーバは研究室で管理
    - 汎用WSレベルで実現可能
    - DB環境はオープンソースを利用（PostgreSQL）
  - 公開サーバは研究室の管理から切り離す
    - 汎用的な共通仕様　・・・・　データ管理者の負担軽減
    - 公開サーバへは複数のDBサーバから接続可
    - ある拠点での集中管理も可能とする

データ閲覧システムの構築
（2） データへのアクセス権限の管理

柔軟なデータ公開ポリシーの設定

「閲覧できる／できない」といったレベルでは不十分

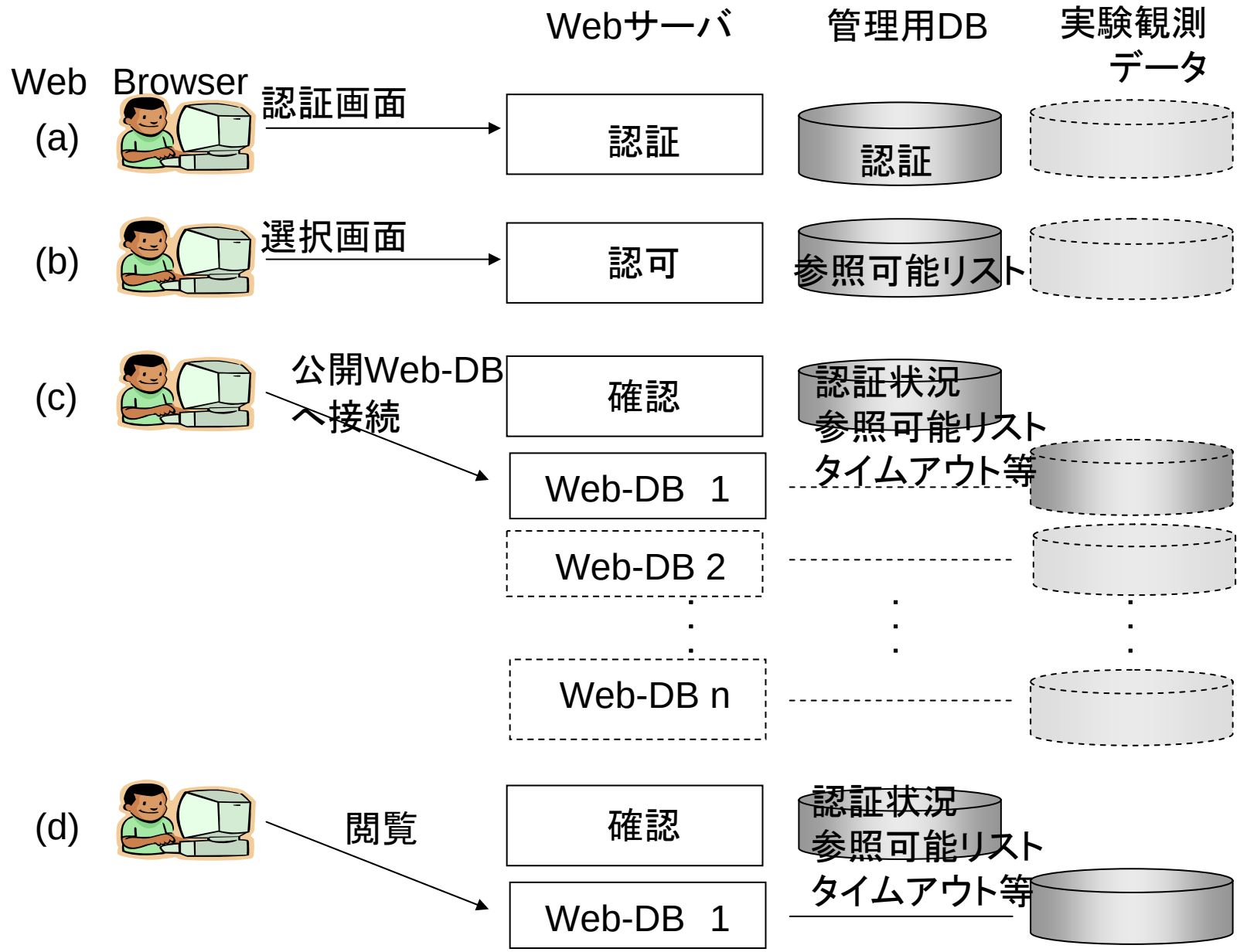
⇒ データ公開レベルをユーザ・グループごとに柔軟に設定

例

- 生データはデータ所有者（取得者）のみ閲覧可能
- 較正済データは共同研究の関係にある研究グループのみ閲覧可
- 観測データのサマリや低解像データは全ユーザが閲覧可
- 特別な研究プロジェクトやキャンペーン観測など，特定データを一部ユーザに限定して閲覧可

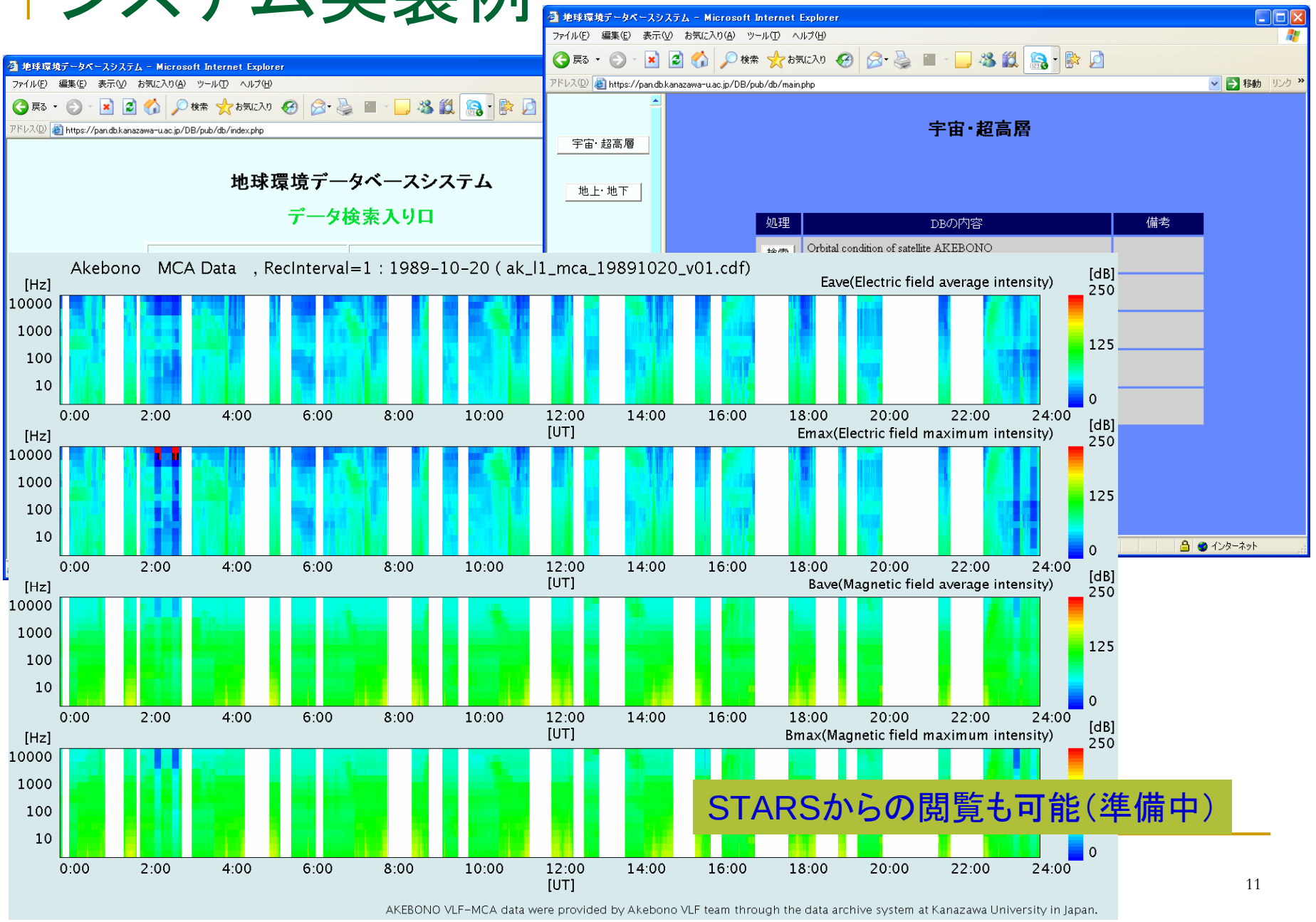
etc.

システム動作の流れ


Webサーバ
管理用DB
実験観測データ
Web Browser
(a)
認証画面
認証
認証
(b)
選択画面
認可
参照可能リスト
(c)
公開Web-DBへ接続
確認
認証状況
参照可能リスト
タイムアウト等
Web-DB 1
Web-DB 2
Web-DB n
(d)
閲覧
確認
認証状況
参照可能リスト
タイムアウト等
Web-DB 1

システム実装例

https://wwwdb01.db.kanazawa-u.ac.jp/DB/



STARSからの閲覧も可能（準備中）

我々に必要なものは何か？

- テラバイトオーダーのデータの蓄積
  - 大量データのDB化と公開システムの整備
  - 研究者に負担がかからないデータ生成・蓄積機構の構築
  - 誰にでもわかりやすいデータフォーマットの定義
  - 適切なデータ公開ポリシーに基づくデータ閲覧機構

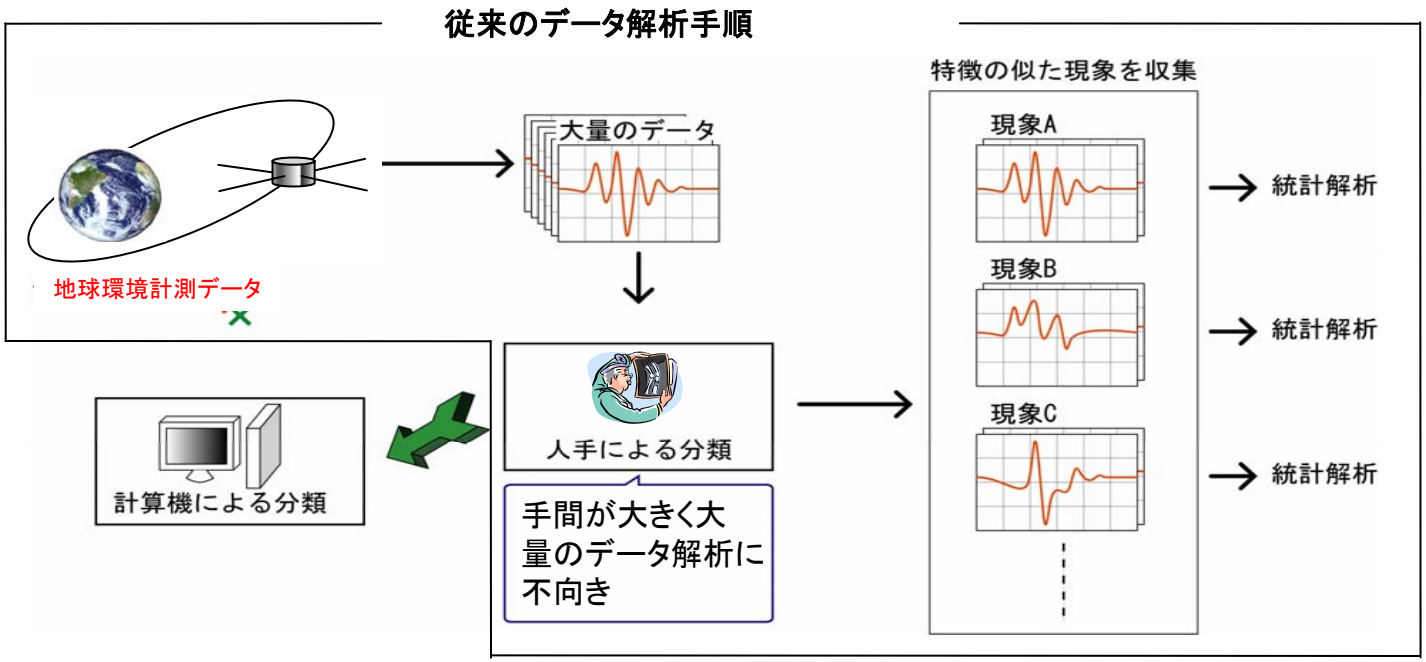
- 所望の観測条件に合致したデータの検索

  → 刻々と変わる観測条件・観測モードから
  適切なデータをどう抽出するか？

  - 適切かつ汎用的なメタデータの定義
  - メタデータ収集・検索機構の充実 ・・・ OAI-PMH、RSS etc.

- データ解析者が所望するデータの抽出
  - 解析者にとって必要なデータと無用なデータの識別

大規模データベースからの発見的情報の自動抽出

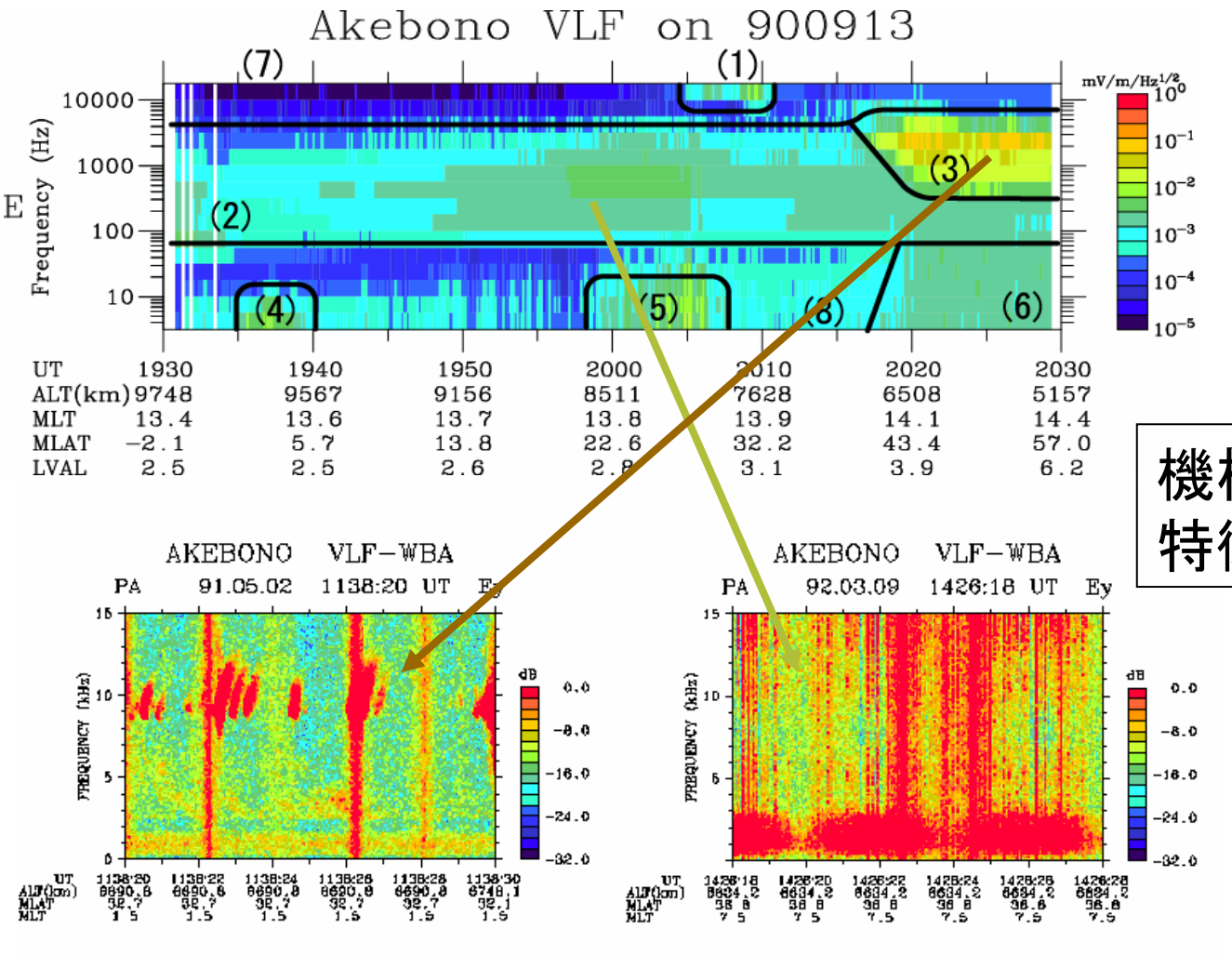


Tbyteオーダーのデータをすべて（専門知識を有する）人間がサーベイし、解析することは事実上不可能！

⇒ データを専門家と同じ抽象的概念に基づき、計算機が自動判別・分類するしくみの提案

蓄えるだけでは成果につながらない！

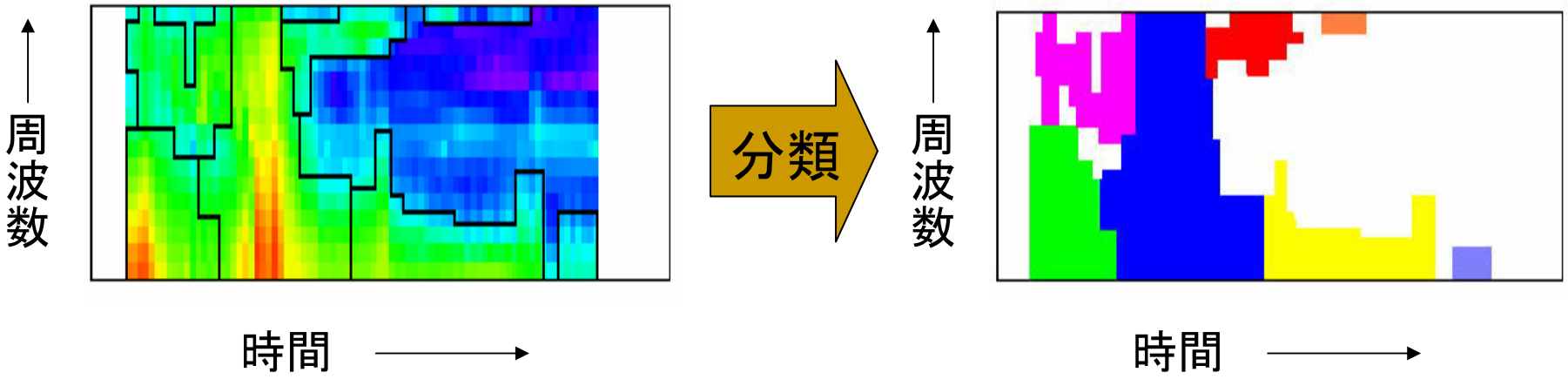
データの特徴の自動認識・分類



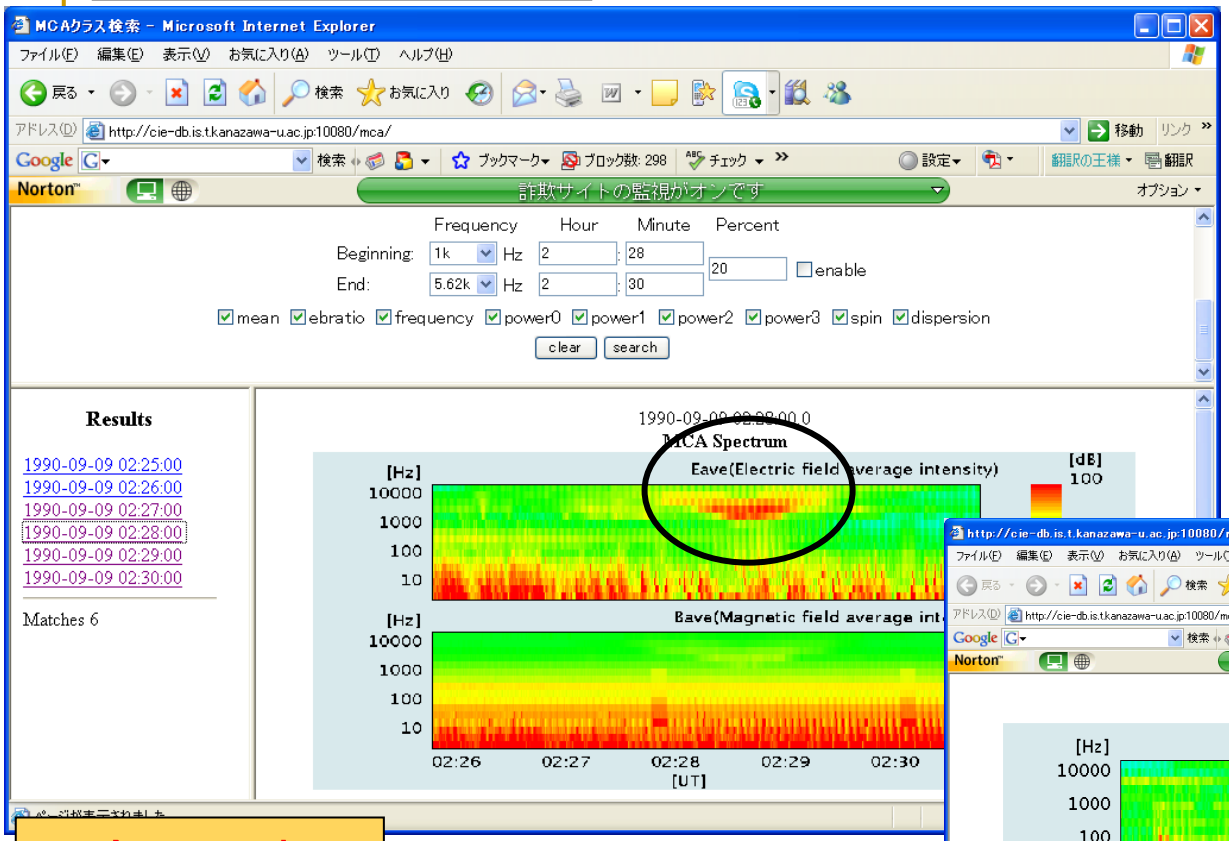
機械的にデータの
特徴を識別したい！

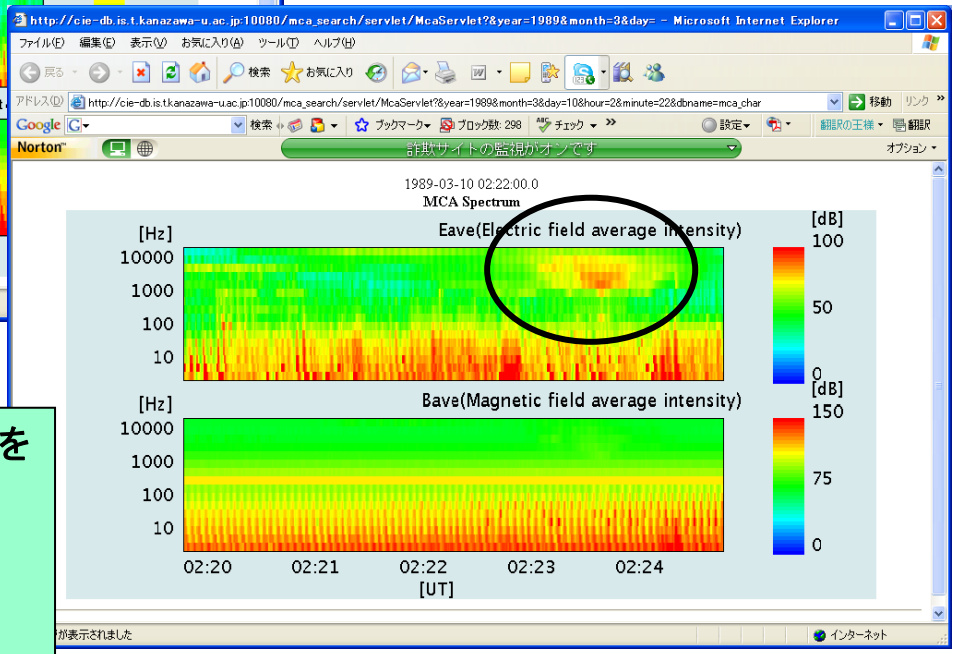
Event Finder の開発

- 研究者が用いる評価基準の定量化・正規化
- データの種類によらないデータ識別アルゴリズムの汎用化
- ブラウザによる分類結果の検索・表示システムの開発
- 興味深いデータを抽出するアルゴリズムの開発



試作システム例





研究開発諸元

- データの**意味・特徴**を客観的指標で表現し、大量データを**計算機の手で分類・体系化**したデータベースの実現

  （**Automatic indexing**）

- 体系化したデータベースから、あいまいな検索語を柔軟に解釈し、**特徴的な未知・発見的**データを検索・抽出

  （**Event finder system**）

まとめ

- **データをためるシステム（入れ物）の作成は比較的容易**
  - 馬力（お金・人・計算機資源）があれば、ある程度解決
- **データ提供者・利用者相互のニーズに合ったシステム設計**
  - 提供者が使わない and/or 使えないシステムは、他の人も使えない
  - 提供者の義務的負担に負うのみでは維持・発展は困難
  - 地球科学系研究者と情報系研究者の連携強化
- **今後の発展に向けて・・・・**
  - システム利用効果のアピール
    DBシステムの威力を我々がフルに活用！