DandDによる
関係形式の高度利用

慶応義塾大学理工学部
横内大介
柴田里程

データの組織化

- 組織化の多様性
  - データベースの構築
  - 解析ソフトウェアによるデータ解析
  - etc

生データはあるが・・・

どのように組織化すればよいのか？

この組織化でよいのか？

・・・・・

あやめデータ

- 測定
  - あやめ 3 品種 Setosa,Versinica,Versicolor, 各50個体について，がく片および花弁の幅，長さを計測

I.D.	がくの長さ	がくの幅	花弁の長さ	花弁の幅	品種
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
…	…	…	…	…	…

データの組織化

- 組織化の多様性

その 1

I.D.	がくの長さ	がくの幅	花弁の長さ	花弁の幅	品種
1	5.1	3.5	1.4	0.2	setosa
…	…	…	…	…	…

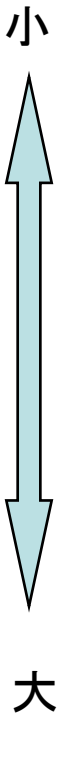
その 2

I.D.	品種	測定部分	長さ	幅
1	Setosa	がく	5.1	3.5
1	Setosa	花弁	1.4	0.2
…	…	…	…	…

その3

I.D.	品種	測定部分	計測項目	計測値
1	Setosa	がく	長さ	5.1
1	Setosa	がく	幅	3.5
1	Setosa	花弁	長さ	1.4
1	Setosa	花弁	幅	0.2
…	…	…	…	…

データ部に含まれる情報



これらの関係形式データの組織化の違いはなにか？

⇒ 関係形式の各行に注目すると

I.D.	品種	がく片の長さ	がく片の幅	花弁の幅	花弁の長さ

1 行は「あやめの花」をあらわしている

I.D.	品種	測定部分	幅	長さ

「あやめの花の部位」

(測定部分に幅と長さある部位のみ)

I.D.	品種	測定部分	計測項目	計測値

「あやめの花の部位の一回の計測」

関係形式の各行の裏にあるオブジェクトを
Target Object とよぶ

Attribute(属性)　　Profile(外見，見かけ)

I.D.	品種	がく片の長さ	がく片の幅	花弁の幅	花弁の長さ

「あやめの花」

I.D.	品種	測定部分	幅	長さ

「あやめの花の部位」

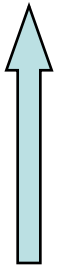
I.D.	品種	測定部分	計測項目	計測値

「あやめの花の部位の計測」

下流

データに対する見方を反映

（データ解析を意識した組織化）



データの取得現場

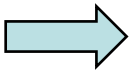
（データ更新の利便性を意識した組織化）

上流

データの変容

- あやめデータ

あやめの花の部位
の計測



あやめの花の部位

あやめの花

・データの組織化，モデル化のヒントを示す Target Object

・データに対する見方，その構造の変化 → データの変容

データの解析，モデル化の過程で頻繁に生じる
データの変容をいかにサポートすべきか

データの変容(1)

- データの編集作業について考える
  - 表計算ソフトウェア
    - 利点
      - 視覚的な編集操作
    - 欠点
      - 明確なデータ型が存在しない
      - データ以外のイレギュラーな記述 の許容
  - S-PLUS，Rのような本格的な統計解析ソフトウェア
    - 利点
      - データの編集作業に必要十分な機能
      - 明確なデータ型とデータ構造
    - 欠点
      - 高度な関数，ソフトウェアの知識
      - 考えながらの作業

データの変容(2)

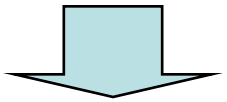
- データの変容の記録
  - 従来の環境では
    - 時間の経過とともに，解析，モデル化した人間でさえその詳細を思い出すことは困難

つまり，

・目的の定まったデータの編集作業

・データの変容の記録

に対する適切なサポートが必要になりつつある．



DandDデータ統合環境

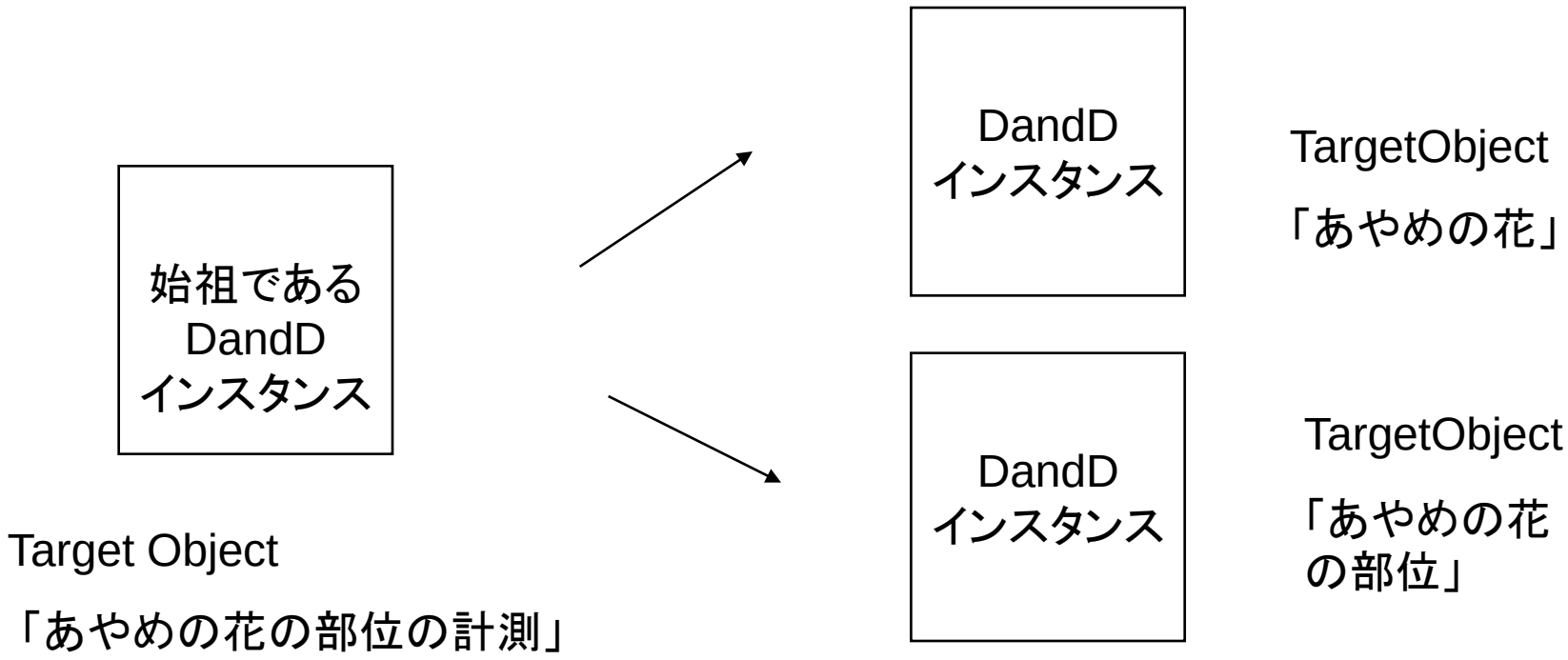
DandD環境によるデータの組織化

- XML文書（DandDインスタンス生成）によるデータの組織化と変容

データファイルからのDandDインスタンス生成

RDBからの DandD インスタンス生成

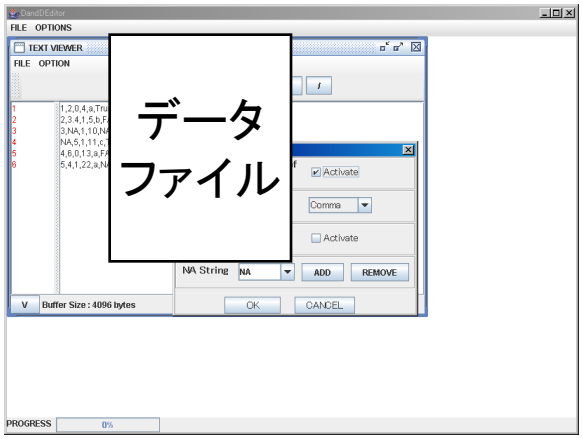
DandDインスタンスのネットワーク



十分な属性情報を与えつつ，新たなDandDインスタンスを作成しながら，データの変容を表していく.

データ
ファイル

DandDEditor



RDB
データ

DandD
インスタンス

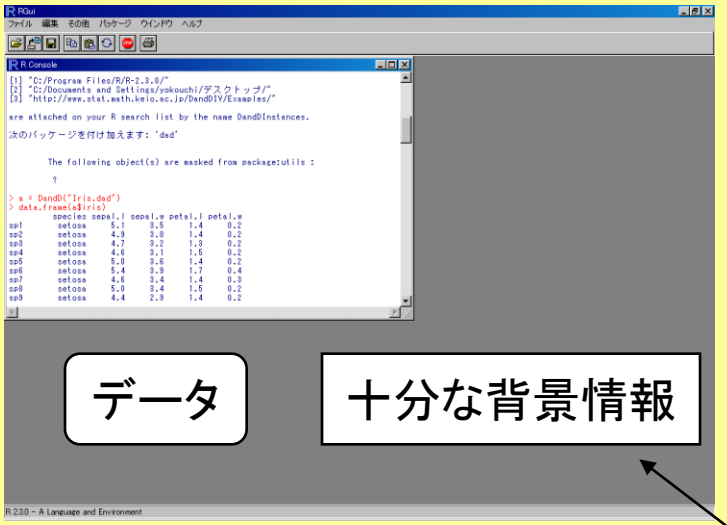
DandDインスタンスが出来上がれば，以後ユーザーはいつでも容易にデータとその十分な背景情報を取得できる

ユーザー（クライエント）

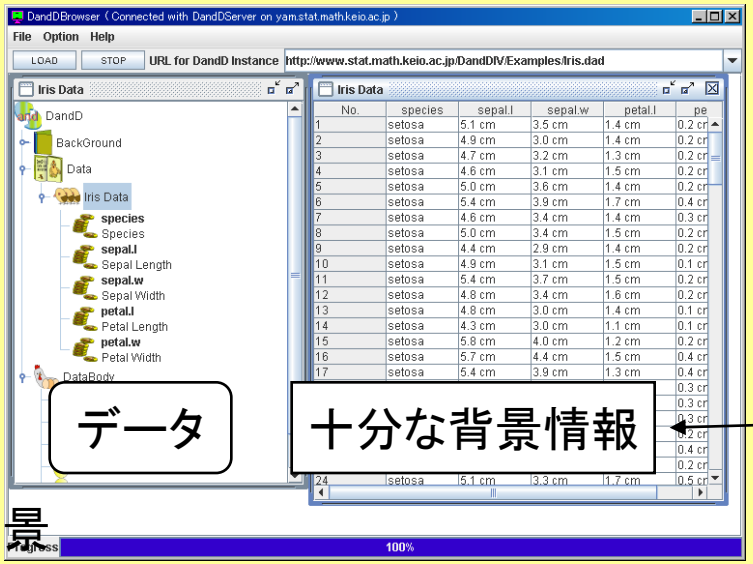
ブラウジング

データ解析

モデル化



DandDR



DandDBrowser

データと背景情報のブラウジング

RDB

データ

DandD インスタンス

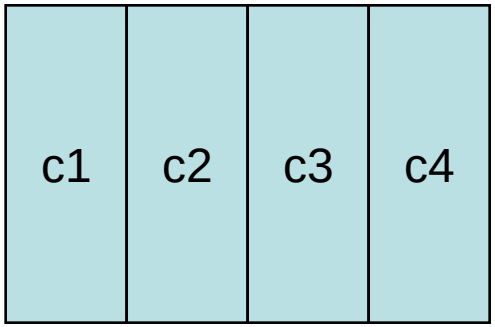
DandD インスタンス

DandDServer

DandDインスタンス における関係形式の表現

```
<Data>
<Relation Columns="c1 c2 c3 c4">
</Data>
<DataBody>
<DataVector Id="c1"/> (in DB1)
<DataVector Id="c2"/> (in DB1)
<DataVector Id="c4"/> (in DB1)
<DataVector Id="c4"/> (in DB1)
</DataBody>
```

DB1



各データベクトルには，RDBへ所在や SQL によるデータの取得手続きなどの情報が埋め込まれている.

関係形式データの変容

<Data>

<Relation Columns="v1 v2"/>

</Data>

<DataBody>

<DataVector Id="v1" RefId="C1 C2"/>

<DataVector Id="C1"/> (in DB1)

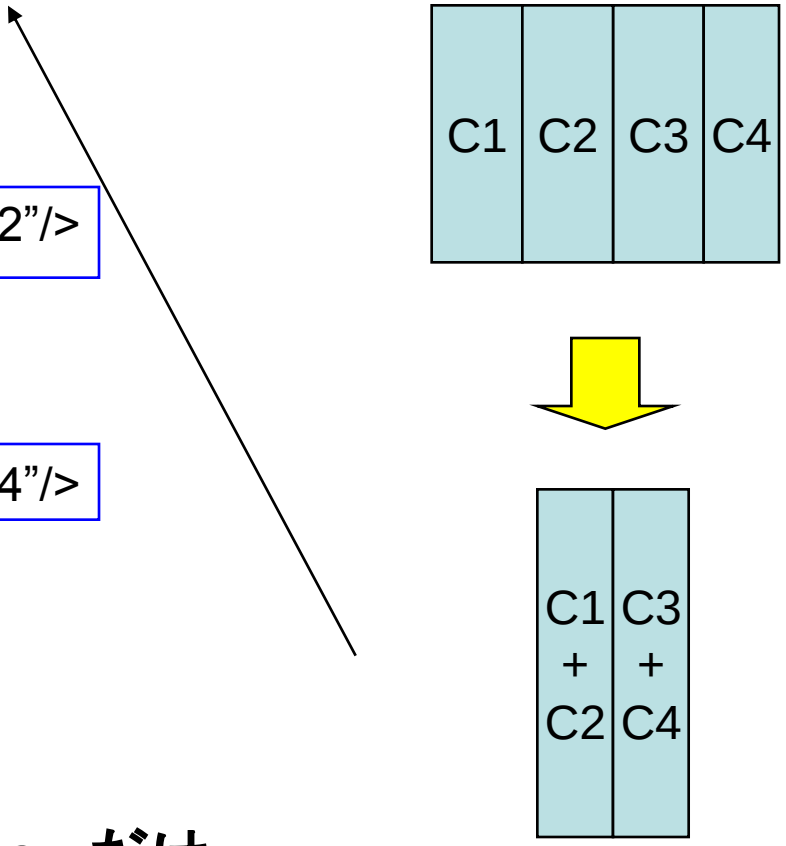
<DataVector Id="C2"> (in DB1)

<DataVector Id="v2" RefId="C3 C4"/>

<DataVector Id="C3"/> (in DB1)

<DataVector Id="C4"> (in DB1)

<DataBody>



元データはそのままに，Viewだけを変容することが可能

<Data>

<Relation Columns="C1 C2 C3 C4"/>

</Data>

<DataBody>

<DataVector Id="C1"/> (in DB1)

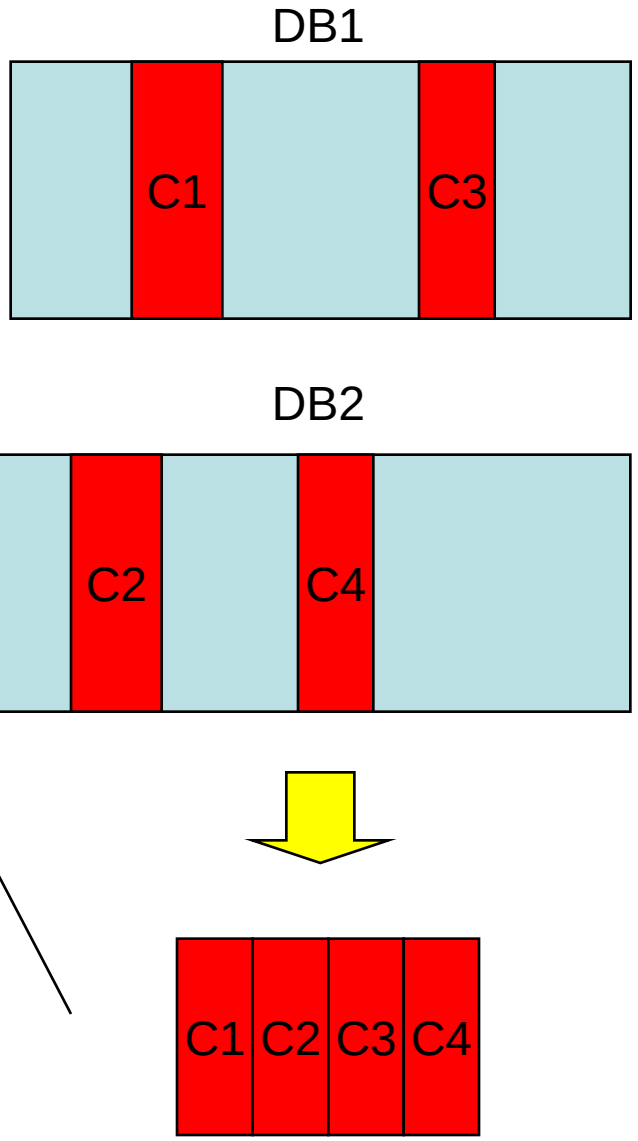
<DataVector Id="C2"> (in DB2)

<DataVector Id="C3"/> (in DB1)

<DataVector Id="C4"> (in DB2)

<DataBody>

複数の関係形式データを1つの関係形式に統合することも可能.



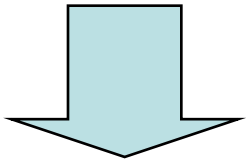
SQLによる関係形式データの変容

I.D.	品種	測定部分	計測項目	計測値

関係形式 あやめの計測

```
<Query Id="Sepal.Length">
    SELECT 計測値 FROM あやめの計測
</Query>
```

関係形式における計測値カラムの値の並び



```
<Query Id="Sepal.Length">
    SELECT 計測値 FROM あやめの計測  WHERE 品種='Setosa'
</Query>
```

品種が Setosa である計測値の並び

関係形式の属性情報

- Target Object
- 特別な意味をもつカラムの組
  - Hypo-Relation
    - 基数系
    - 座標系
    - AttributeとProfile
- 複数の関係形式の間に存在するデータベクトル同士の関係
  - Hyper-Relation
    - SharedValue
      - 変量としての同一性
      - 値の範囲，意味は等しい
      - いわゆる関係形式の外部キーはこの一種
    - CommonMeasurement
      - スケールの共有

DandDデータ統合環境

- DandDによるデータの組織化，変容
  - 十分な背景情報と変量の属性情報
  - DandDEditorによるデータ編集記録（ロギング）
  - 形式的な変容記述
    - 一度RDBに放り込まれたデータには一切手が加えられないので，データの変容のトレースは容易
  - DandDEditorによる視覚的なデータ操作
    - DandDに関する高度な知識は不要
- 「とりあえず」からの卒業
  - 目的を明確にしたソフトウェアの設計，開発
  - アドホックな機能の追加を回避

DandDプロジェクトのホームページ

http://www.stat.math.keio.ac.jp/DandDIV/index.ja.html