Hitotsubashi University
Graduate School of
International Corporate Strategy

# データサイエンス実践の支援環境 TRAD

横内大介（一橋大・ICS）
柴田里程（慶應大・理工）

# アウトライン

- 先行研究
  - DandD，InterDatabase
  - TextilePlot
- TRADの紹介
  - TAD
  - TRAD
- TRADにおけるデータ操作機能について
  - 正規化と非正規化
    - データマネージメントの観点
    - データ解析の観点
- 今後の課題

# TRADとは

- TAD＋R=TRAD
  - TAD＝TextilePlot and DandD
  - R
- フリーウェア
  - ホームページからダウンロード可能
- Javaによる実装
  - JRE 1.8.0_70 以上で動作
- 主要な3つのコンポーネント
  - データと背景情報，属性情報の一体化およびInterDatabaseを実現するDandD
  - 高次元データをそのままながめることができるTextilePlot
  - データサイエンス実践の基本的な道具であるR

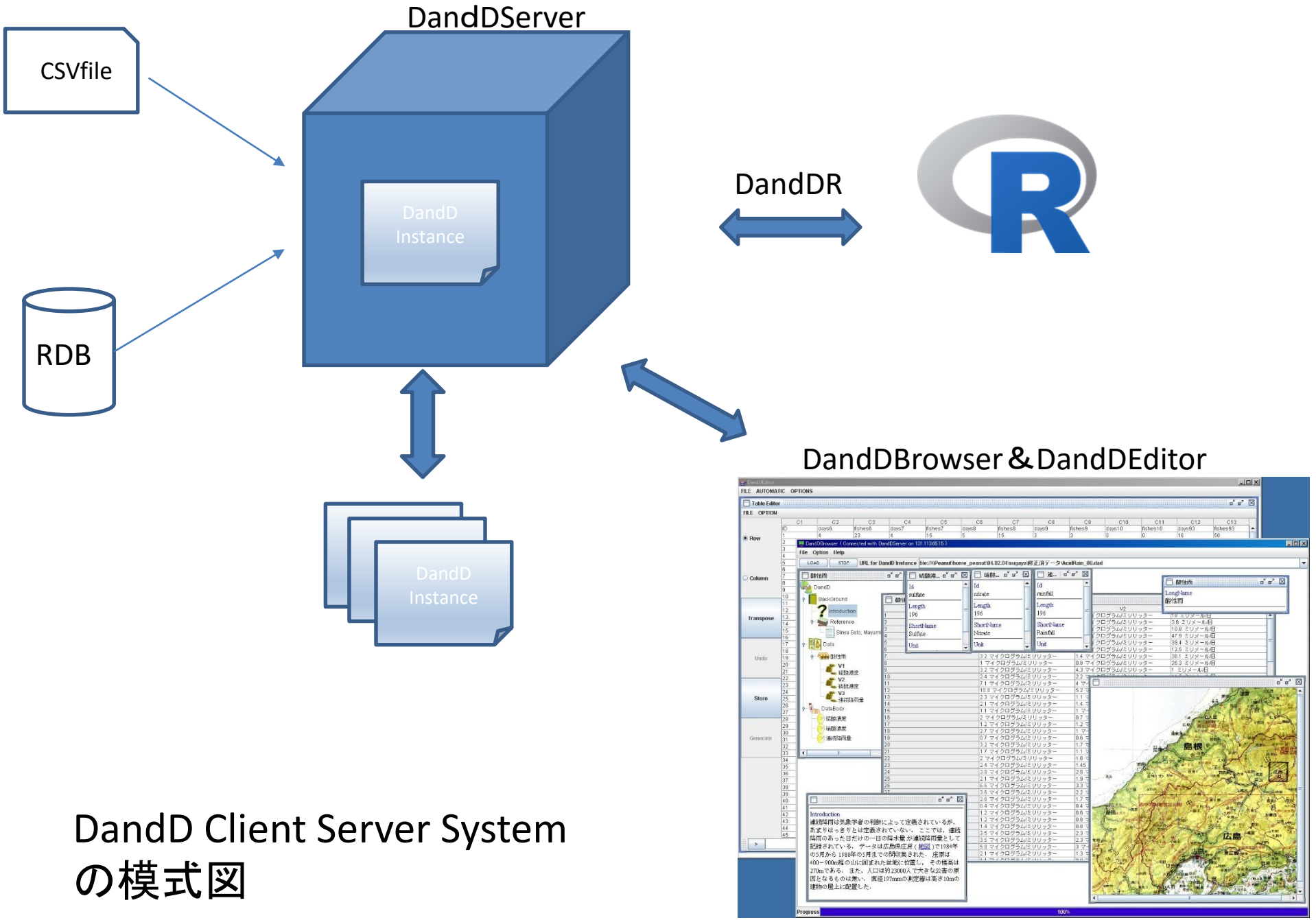
データサイエンスコンソーシアムのHP　http://datascience.jp/

# DandD

- DandD（Data and Description）とは
  - DandD rule
    - データとその背景情報，属性情報を一体化するための記述ルール
    - データをData Vector単位に分解し，各DataVector に十分な情報を与えたうえで再組織化するのが大きな特徴
    - XML(eXtensible Markup Language)による実装
  - DandD Instance
    - DandDルールに沿って記述されたXML文書のこと
    - 2つのデータベクトルの格納形式
      - Internal DataVector：データをXML文書内に直接記述
      - External DataVector：実体の代わりに所在，アクセス方法，加工方法を記述する
  - DandD Client Server System (Yokouchi and Shibata[2004])
    - DandD Server
      - DandD Instance に関するすべての操作を提供するサーバシステム
    - DandD Client Software
      - DandD Server との通信を通じて DandD Instance 内のデータを操作するユーザーインターフェイス群
      - DandD Editor，DandD Browser，DandDR

# InterDatabase

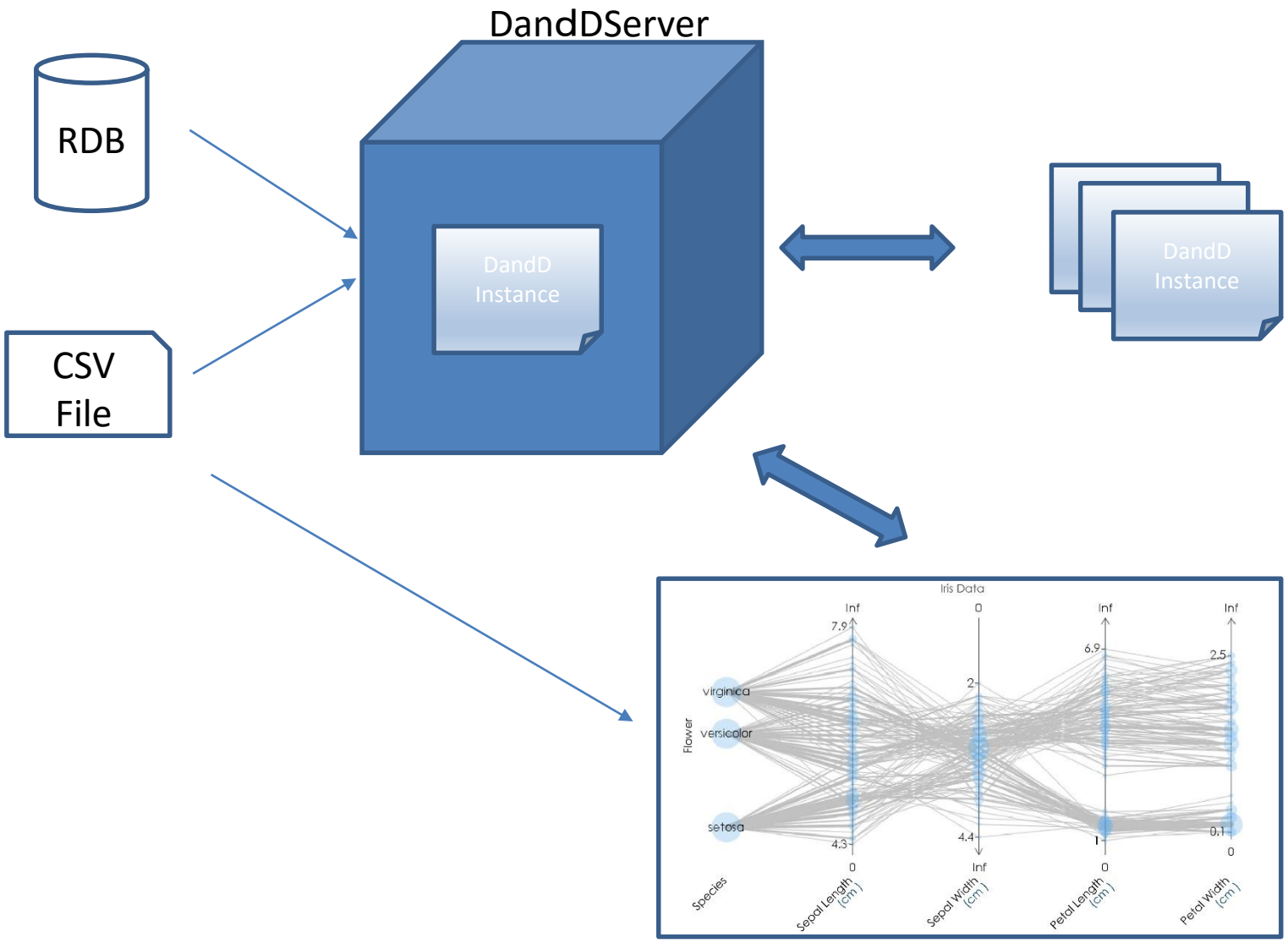
- InterDatabase(横内，柴田[2001])
  - 異種データベースの統合のフレームワーク
- フレームワークの概要
  - ネットワーク上の多様なデータベースをDandD Instance 上のExternal DataVectorとしてそれぞれ記述
  - それらのEternal DataVector を1つのDandD Instance 上で参照し再組織化
  - 実際のデータの取得，加工はDandD Client Server System がDandD Instance にある記述を解釈し実行する
    - データ分析者は，データの取得，加工に煩わされない




DandD Client Server System
の模式図

# TextilePlot

- 平衡座標プロットの一種
- 水平性基準が大きな特徴
  - 各折れ線がなるべく水平線に近くなるよう各座標軸の位置と尺度を決定する基準
- 豊かなグラフィカル表現
  - 数値変量以外に類別変量や欠損の表現も可能
  - データに対して十分な属性情報が与えられている場合はそれらもプロットに反映
    - そのためDandDとの親和性も高い
- ソフトウェアとしても実装済み
  - 熊坂，柴田(2007)によるJavaをもちいた実装
    - 関係式になっている自由欄形式のテキストデータやRDBに対応
    - DandDServerとの通信もできるので，DandD Instance のデータの読み込みにも対応



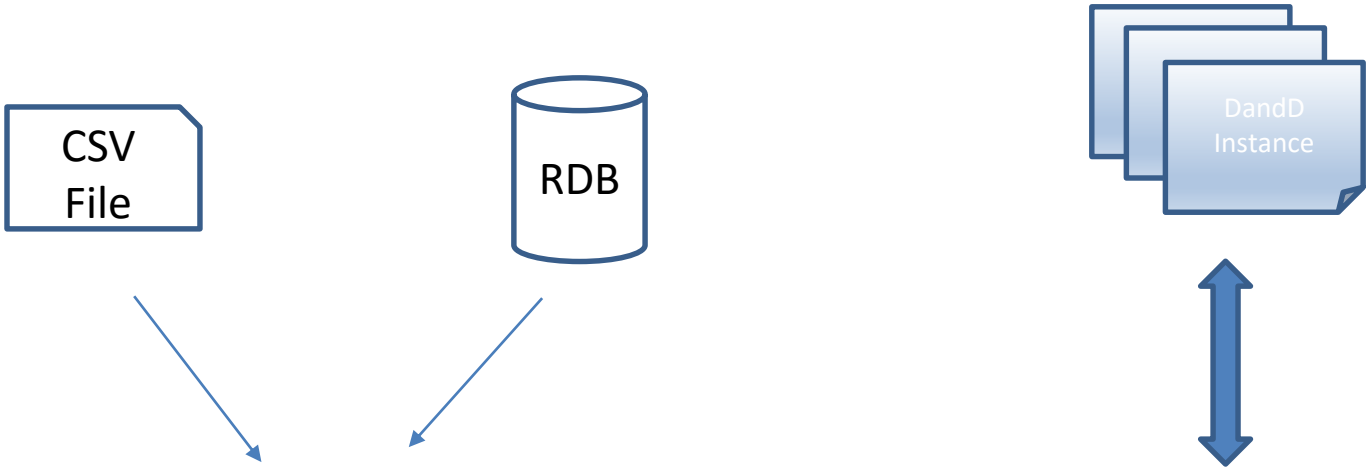

TextilePlot Environemnt

TextilePlot環境の模式図

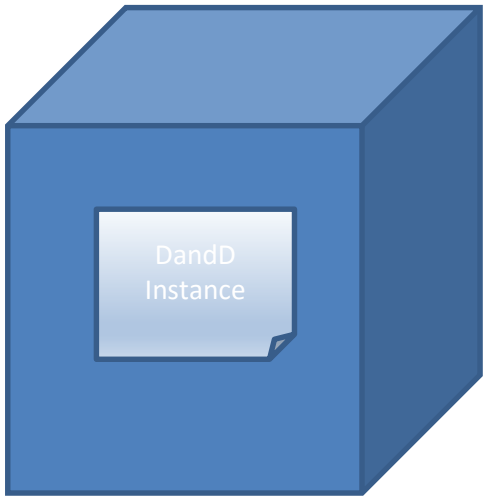
# TAD

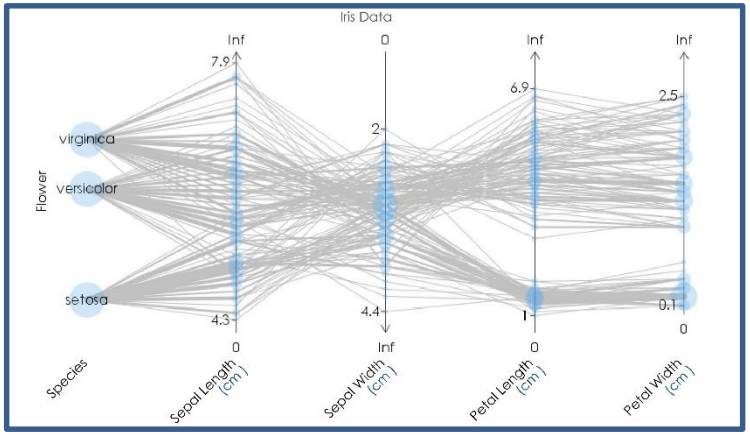
- TADとは
  - TextilePlotとDandDの融合
- 基本設計
  - TextilePlot環境のDataクラスのオブジェクトの初期化をDandD Instance の生成に置き換える.
  - TextilePlot環境上で行われた操作はすべてDandD Instance に記述する.
- 実装方針
  - DandDServerをJavaのクラスライブラリ化し，ビルトインした.
    - TextilePlot環境はJavaで作られているので，Dataオブジェクトの初期化が容易に実現できる.
    - 通信のオーバーヘッドをなくし，処理の高速化が図れる.

# TAD の模式図



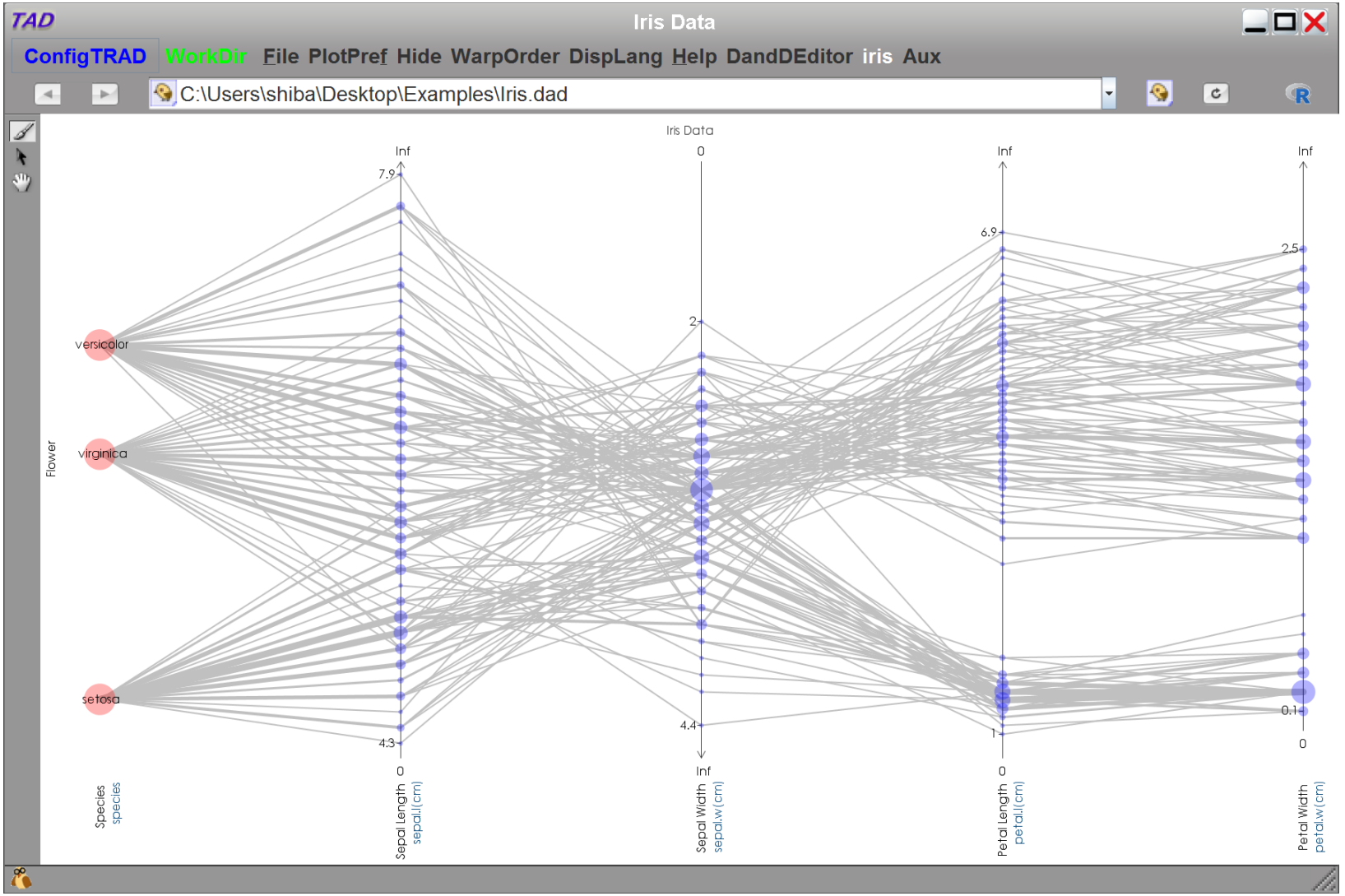

DandDServerをJavaのクラスに再実装



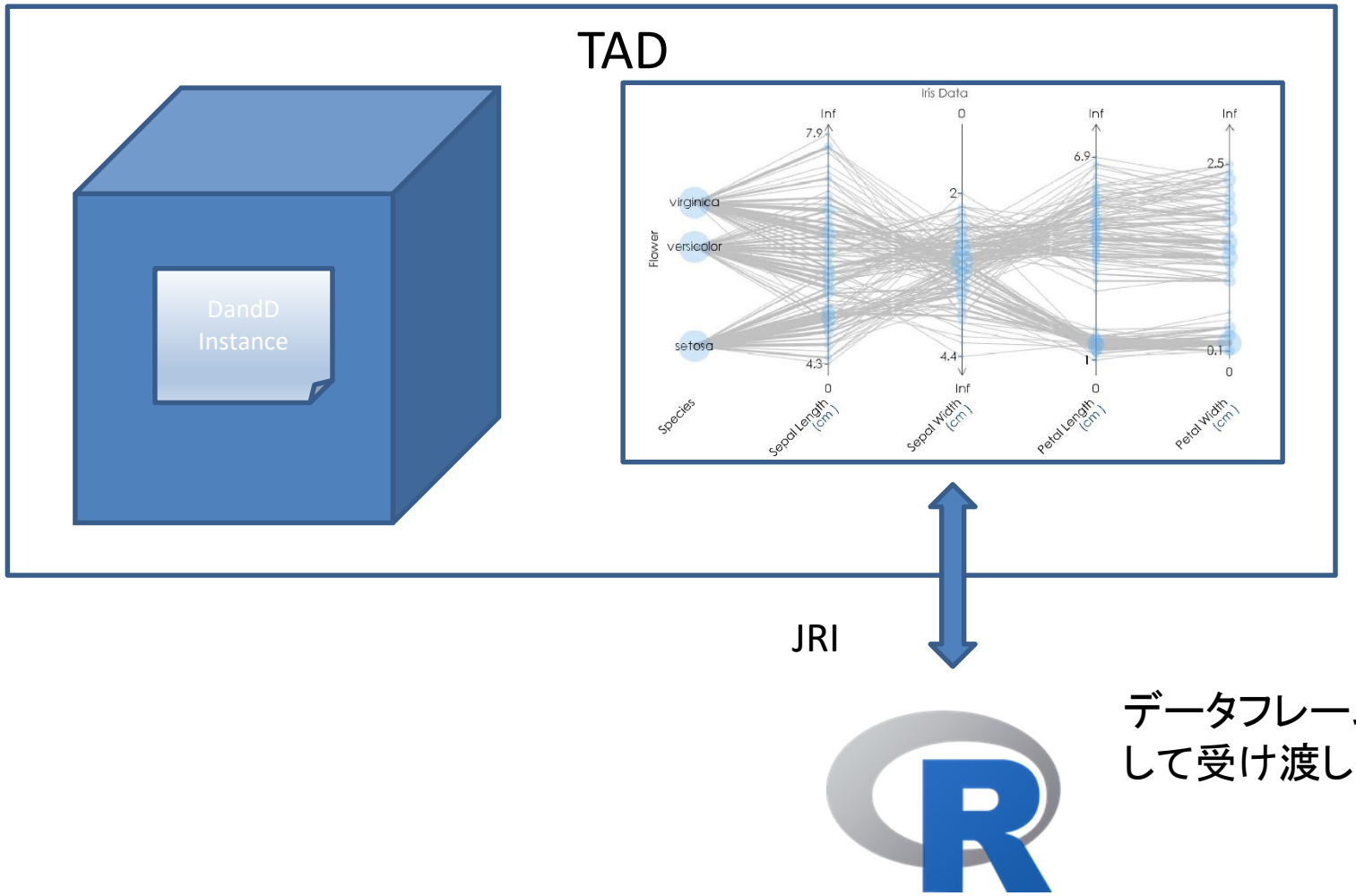

TAD

# TAD によるフィッシャーのアイリスデータの表示



# TADからTRADへ

- TADに不足している重要な機能
  1. 本格的なデータ分析
  2. 十分なデータ操作
- 解決方法
  - 1. について
    - 自前で実装するか，既存の仕掛けを活用するか
      - Rとの連携を選択
        - » JRI (Java/R Interface) パッケージの活用
  - 2. について
    - データ分析に適したデータ構造とはなにか
      - Data Cleaning の形式化
        - » 非正規なCSVファイルへの対応
        - » Data tidying：NormaliseとDenormalise



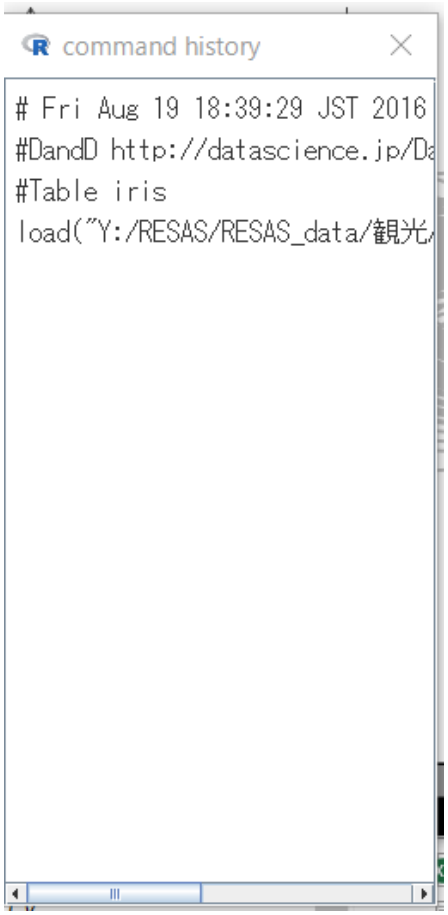

データフレーム化
して受け渡しをする

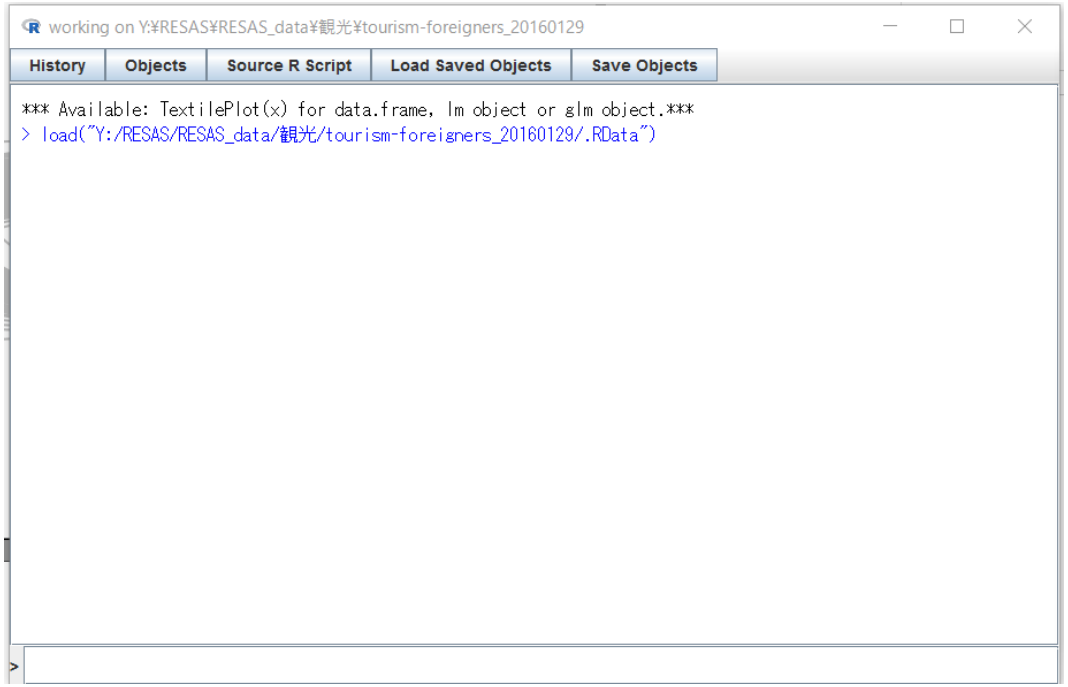
1. の機能に対する現時点での対応

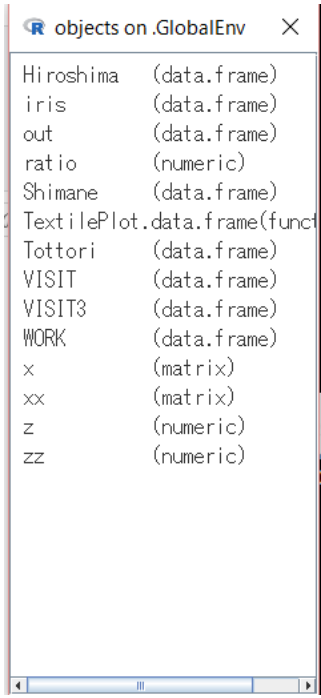
History panel

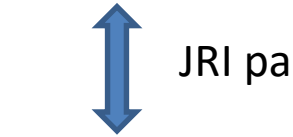
Original JAVA interface to R

Object panel



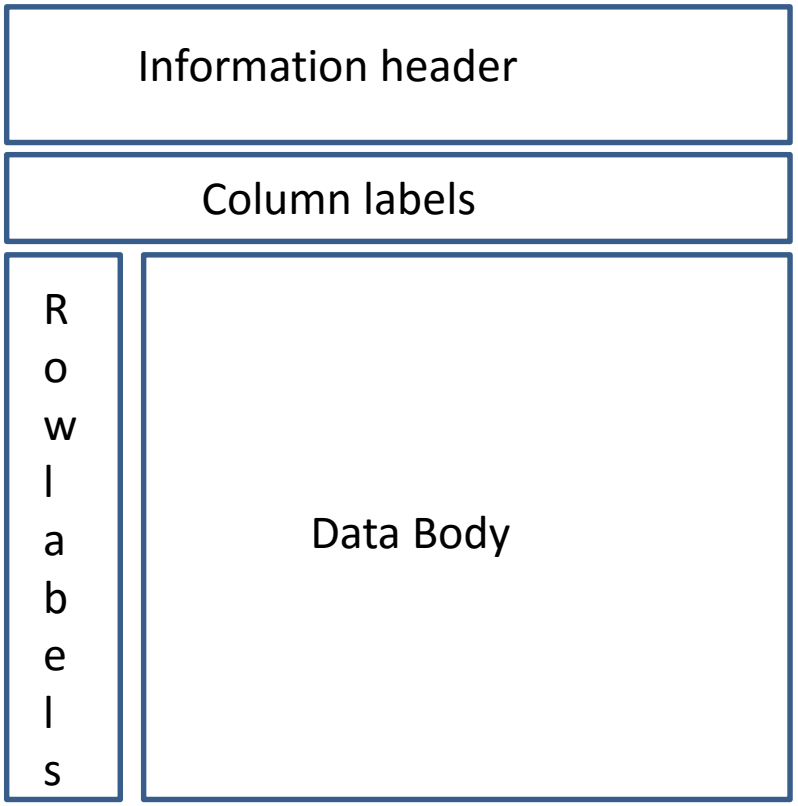
JRI package

TADから呼び出したRへの
インターフェイス

# TRADに実装するデータ操作

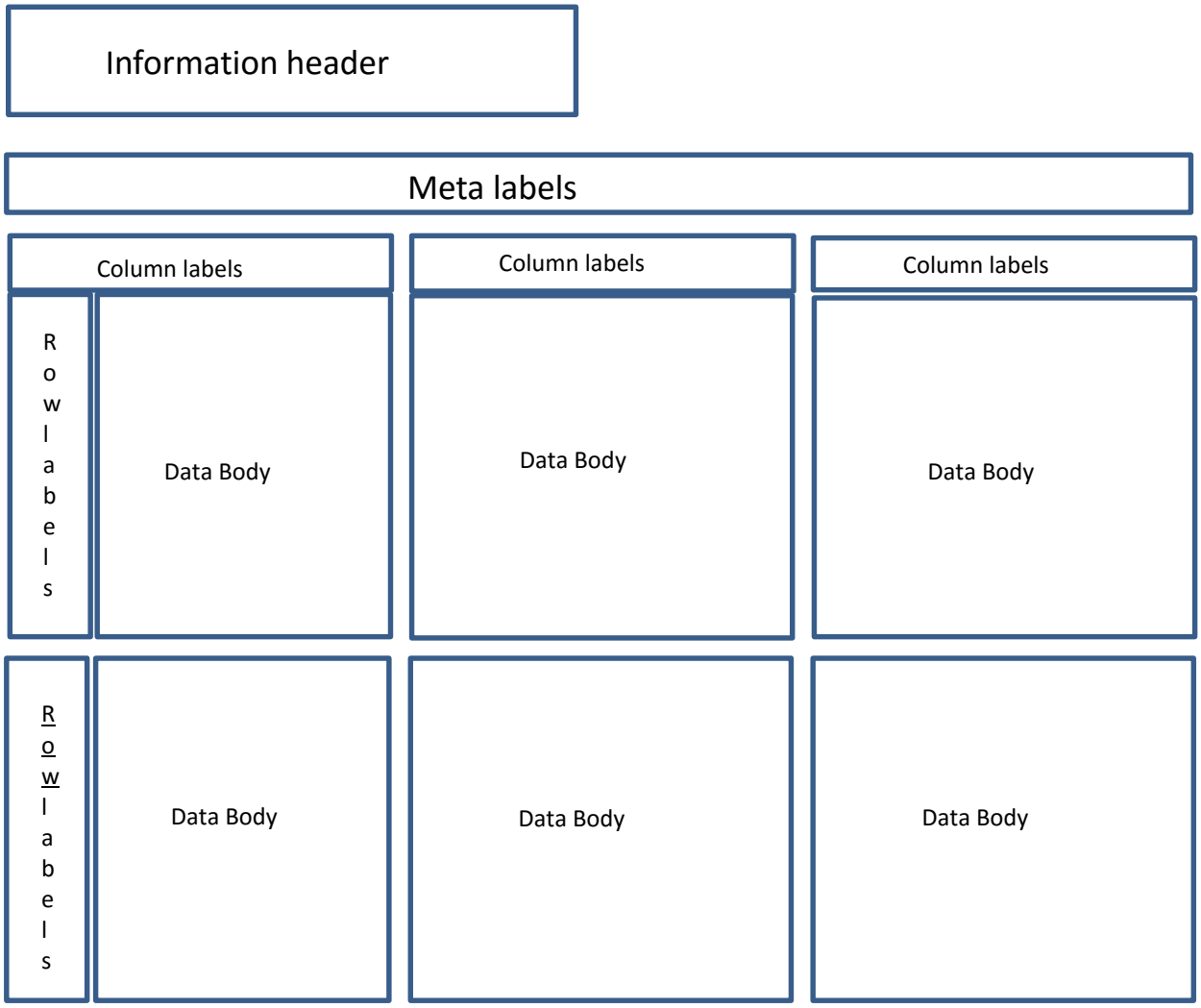
- 2つの機能
  - データテーブルになっていないデータをデータテーブルに直す
    - 現段階では，公的なデータでもよく存在する非正規なCSVデータをテーブルに直す機能を実装
  - テーブルを別のテーブルに直す
    - E. F. Codd の正規化の理論が参考になる．しかし，それをサポートするだけでよいのか？？

# A CSV file stylised



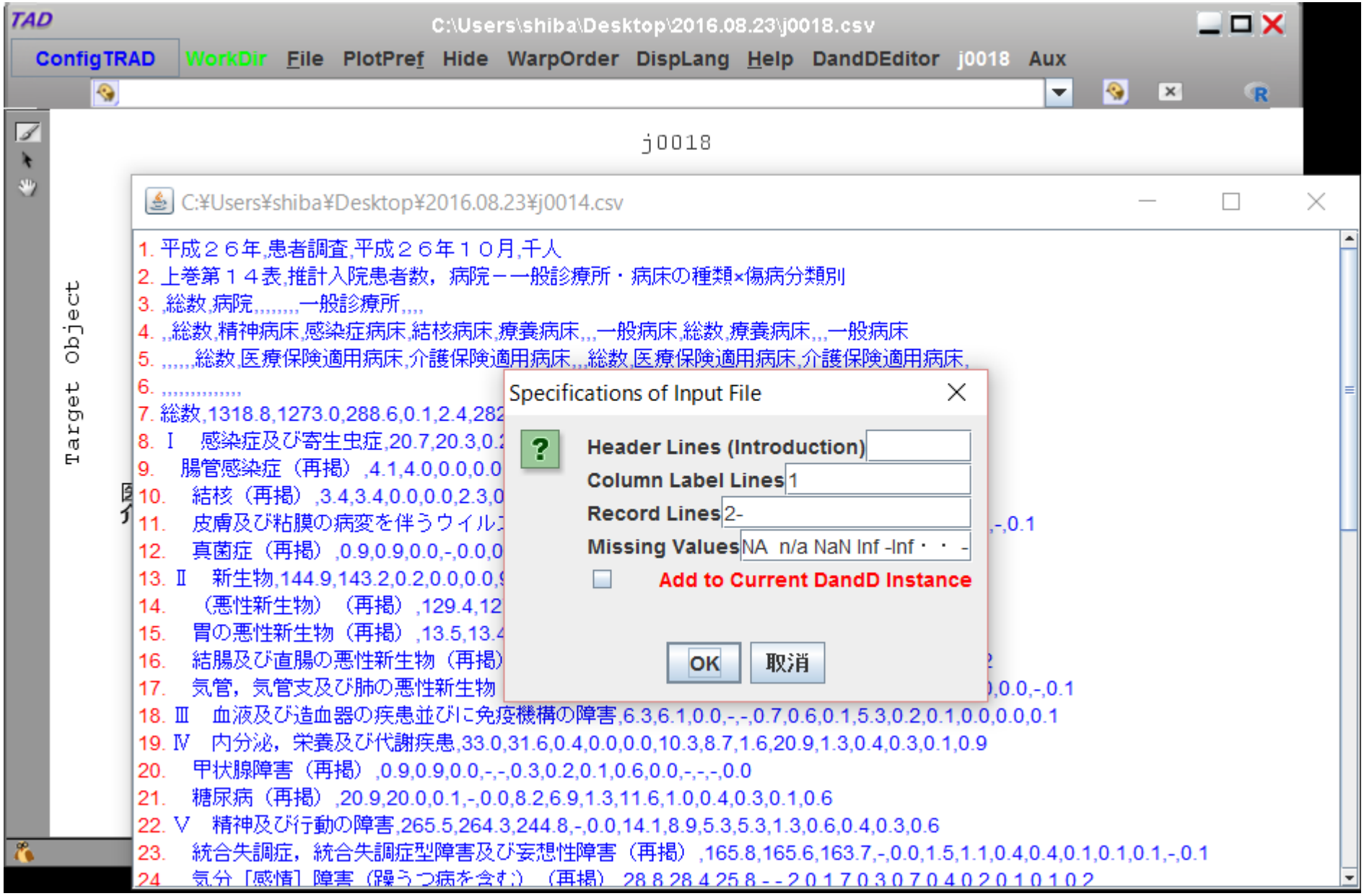
# TADに搭載されているCSV向け機能

- Information header
  - Explanationとして保持する
- Column labels
  - 構造化された複数行にわたるラベルにも対応
  - 区切りとしては空白を想定
- Row labels
  - Column labels と同等のサポート
- Data Body
  - 欠損の再表現
  - 指数表現，quoted or unquoted ,…
  - 転置

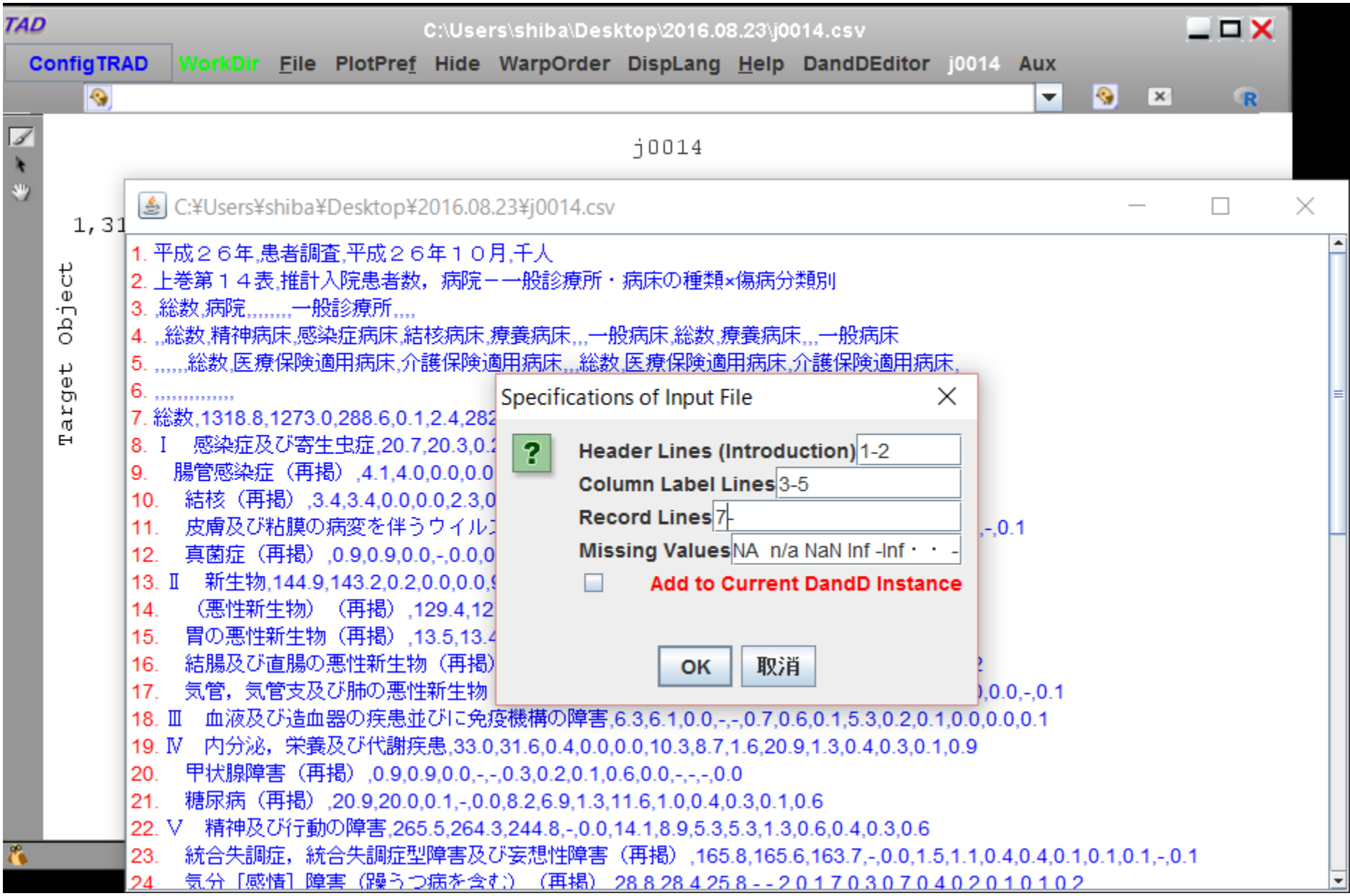



Ex. Open Data : MHL Patient Survey Data

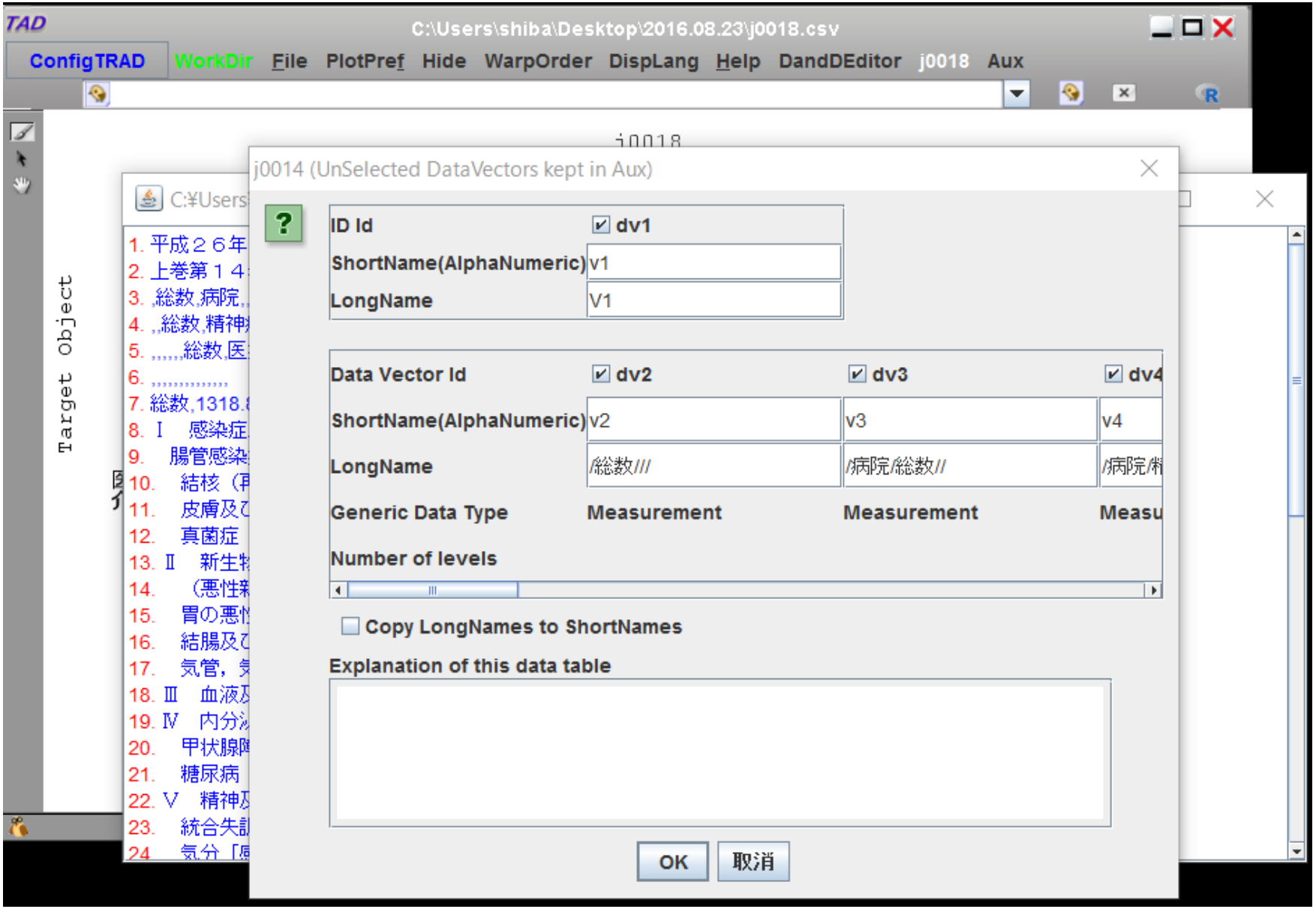
Initial window when a CSV file is drag&dropped.



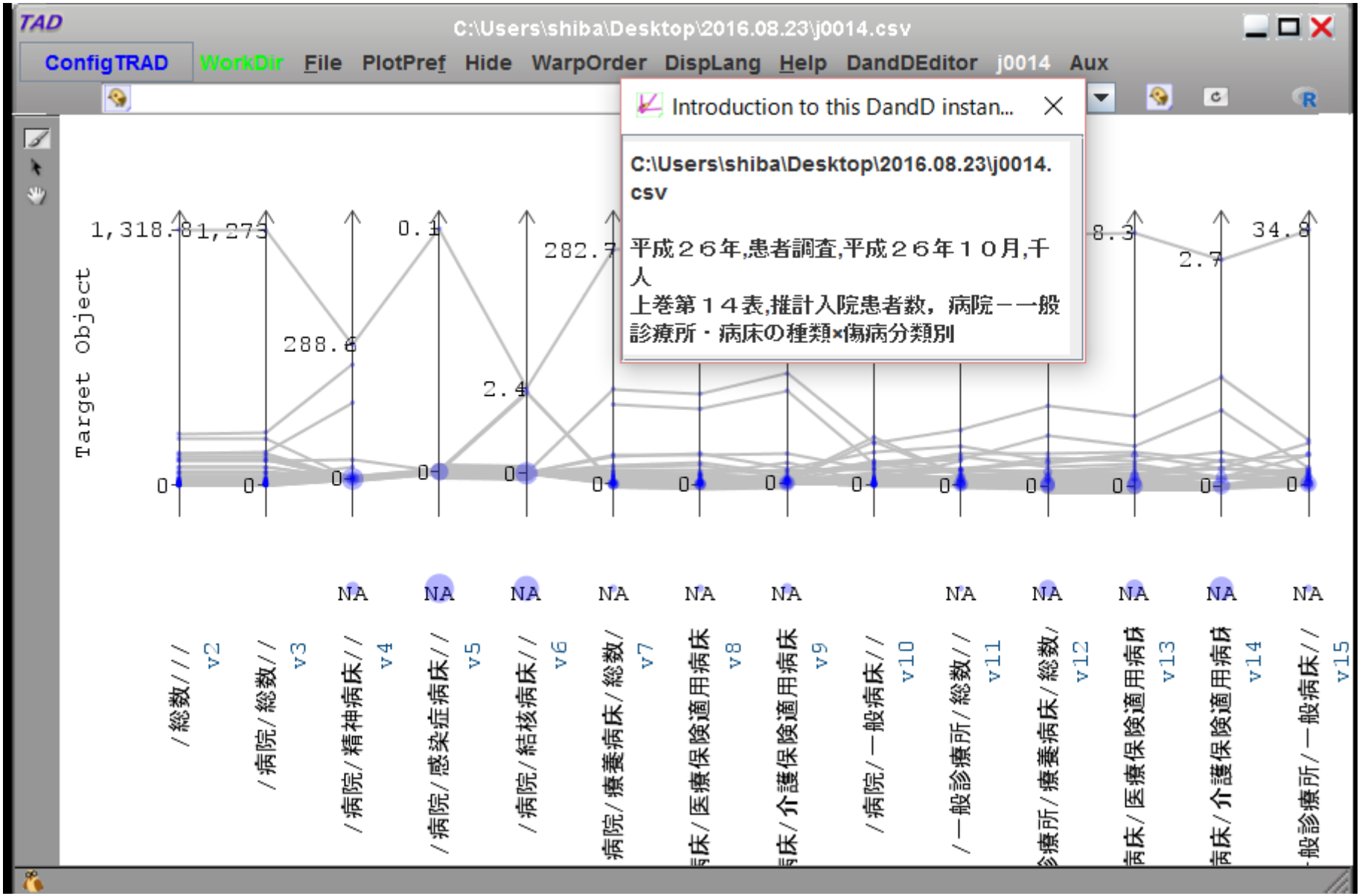
# Specifications



# After reading the file



# Header information



# テーブルからテーブルへの変換

- 関係形式の正規形とデータ解析のための
  データの正規形は異なるものである

アイリスデータの2つの組織化.

| Species | Sepal.length | Sepal.width | Petal.length | Petal.width |
|---|---|---|---|---|
| setosa | 4.0 | 3.8 | 2.5 | 3.0 |
| … | … | … | … | … |

| ID | Species | Measured | Measurement | Value |
|---|---|---|---|---|
| 1 | setosa | sepal | length | 4.0 |
| 1 | setosa | sepal | width | 3.8 |
| … | … | … | … | … |

新しい観測項目（thickenss）が追加された場合

| Species | Sepal.length | Sepal.width | Petal.length | Petal.width |
|---|---|---|---|---|
| setosa | 5.1 | 3.5 | 1.4 | 0.2 |
| setosa | 4.9 | 3.0 | 1.4 | 0.2 |
| … | … | … | … | … |

| Species | Sepal.length | Sepal.width | Petal.length | Petal.width | Sepal .thickness |
|---|---|---|---|---|---|
| setosa | 5.1 | 3.5 | 1.4 | 0.2 | NA |
| setosa | 4.9 | 3.0 | 1.4 | 0.2 | NA |
| … | … | … | … | … | … |
| setosa | 5.0 | 3.6 | 1.5 | 0.4 | 0.04 |
| Setosa | 4.6 | 3.1 | 1.4 | 0.2 | 0.08 |

データベースマネージメントでは不都合な「列の追加」と無駄なスペースである欠損が発生するので非正規形として扱われる.

| ID | Species | Measured | Measurement | Value |
|---|---|---|---|---|
| 1 | setosa | sepal | length | 4.0 |
| 1 | setosa | sepal | width | 3.8 |
| … | … | … | … | … |

| ID | Species | Measured | Measurement | Value |
|---|---|---|---|---|
| 1 | setosa | sepal | length | 4.0 |
| 1 | setosa | sepal | width | 3.8 |
| … | … | … | … | … |
| 151 | setosa | petal | length | 1.5 |
| 151 | setosa | petal | width | 0.4 |
| 151 | setosa | sepal | thickness | 0.04 |
| … | … | … | … | … |

こちらの形式は欠損や列の追加という不都合が起きないので
データベースマネージメントの立場では好都合

# 解析のためのデータの組織化

- データ分析の観点からは？
  - アイリスの分類を説明するモデルの構築が目的ならば
    - 個体はアイリスに対応している必要がある
    - DandDルールにおけるTargetObjectを定める作業に相当
      - TargetObject＝iris or part of iris ?
  - データベースマネージメントでいうところの正規形はデータ分析にとっては非正規形
    - つまり正規形を非正規形に変換する機能が必要

| Species | Sepal.length | Sepal.width | Petal.length | Petal.width |
|---|---|---|---|---|
| setosa | 4.0 | 3.8 | 2.5 | 3.0 |
| … | … | … | … | … |

○

| ID | Species | Measured | Measurement | Value |
|---|---|---|---|---|
| 1 | setosa | sepal | length | 4.0 |
| 1 | setosa | sepal | width | 3.8 |
| … | … | … | … | … |

×

# テーブルからテーブルへ

- データベースマネージメントにとっての非正規形がデータベースの正規形かもしれない
  - 例：金融の生データの多くは，データベースエンジニアが正規化を意識して加工，保存している
- 正規化テーブル間
  - 第1正規形から第5正規形の間の変換
    - SQL等による機能と同等の操作をサポートすればよい
- 非正規テーブルと正規テーブル
  - アイリスのケースのようなデータのラベル化のサポート

# Wickhamの Data Tidying を例に

| Species | Sepal.length | Sepal.width | Petal.length | Petal.width |
|---|---|---|---|---|
| setosa | 4.0 | 3.8 | 2.5 | 3.0 |
| | | | | |
| | | | | |

Gather (+ add ID)

Spread

| ID | species | measurement | Value |
|---|---|---|---|
| 1 | setosa | Sepal.length | 4.0 |
| 1 | setosa | Sepal.width | 3.8 |
| 1 | setosa | Petal.length | 2.5 |

Separate

Unite

| ID | species | measured | measurement | Value |
|---|---|---|---|---|
| 1 | setosa | sepal | length | 4.0 |
| 1 | setosa | sepal | width | 3.8 |
| 1 | setosa | petal | length | 2.5 |

# 今後の課題

- DataTidyingのTADへの実装
  - NormailiseとDenormalise機能として実装作業中
    - 片道はすでに実装済み
    - 当面はWickham のRパッケージを併用
      - plyr, tidyr
  - 非正規テーブル間の変換を検討
    - 必要なのか？そもそも存在するのか？

# （参考）TRADの役割

- 多様なデータに対するインターフェイス
  - データの成形
  - データの実体と注釈の分離
- TextilePlotを通したデータの視覚的理解
  - データ全体の様子を把握する
    - 高次元，多記録
    - 異常値
- DandD Instance を活用したデータの変容の記録
  - データファイルからのデータの読み込み
  - データベースシステムからのデータの読み込み
  - Rのデータフレームの読み込み
  - データ操作
    - GUIを通じた操作
    - フィルター
  - Data from R