多ソース多サイトデータの活用

統計数理研究所　力丸佑紀（発表者）
データサイエンスコンソーシアム，慶應義塾大学　柴田里程
統計数理研究所　山下智志

多ソース多サイトデータの活用

- 複数銀行の信用リスクデータ解析

A銀行
・財務データ
・CRITSスコアデータ
・顧客データ
・損失額情報
・与信保全

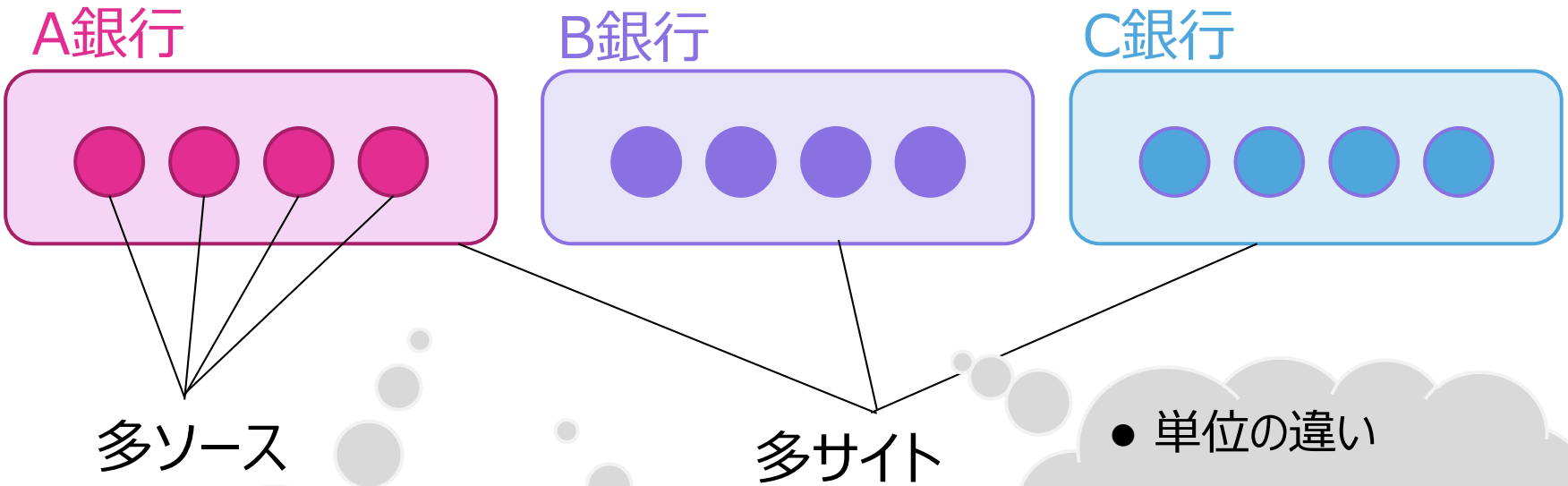
B銀行
・CRITSスコア情報
・期末情報ファイル
・毀損情報ファイル

C銀行
・CRITS情報
・期末情報
・毀損情報

D銀行
・CRITSスコア情報
・コーポレート先
・期末情報
・毀損情報

E銀行
・CRITSデータ
・移管情報
・期末情報
・毀損情報

多ソース多サイト



- 記録の重複

「ID」がIDになっていない

→ 全く同じ記録があった

- 表現の違い

ファイルaでは「2020/03/31」

ファイルbでは「200331」

- 単位の違い

A銀行では「億円」

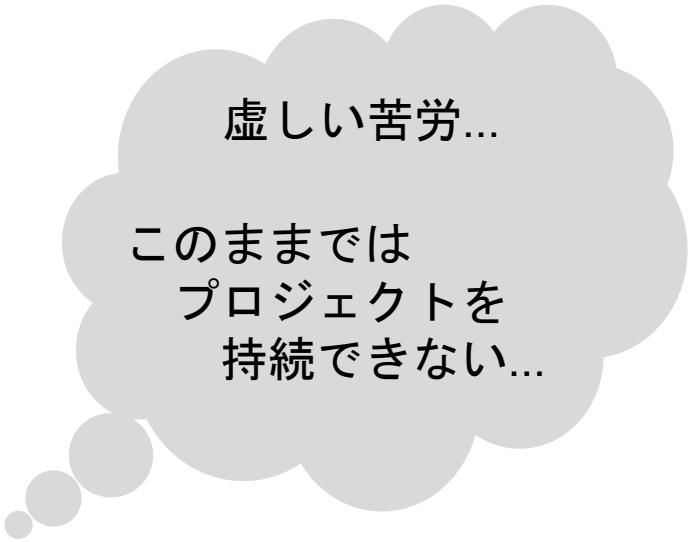
B銀行では「100億円」

多ソース多サイトデータを扱うときの課題

- 不完全性
  - 書式の違い，欠損値の表現の違い，記録の無駄や重複など
- 不統一性
  - 変量の不統一，IDの不統一，単位の不統一，値の表現の違いなど
- 多様性
  - データの多様性，活用目的の多様性など

苦労

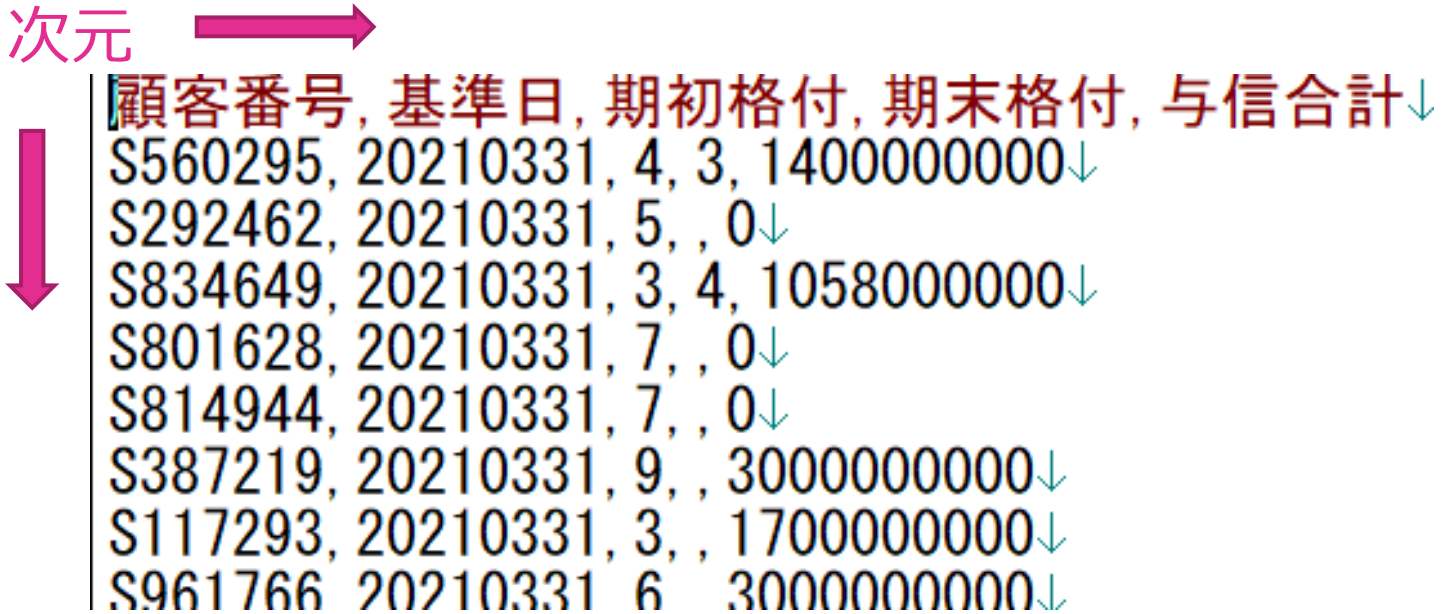
- 手作業で…
- 壮大なプログラムで…
- 業績になりづらい…
- 引継ぎが難しい…



→ 本気でこの課題に取り組むとき

CSVファイル

２次元



プログラム

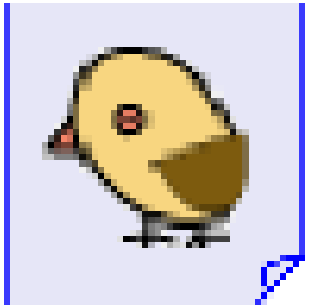
【データの解釈】
- エンコード
- データベクトルの型　・・・

【処理】
- 完全性や統一性などの確保

説明ファイル

【データの説明】
- データの説明
- 変量の説明
- コード対応　・・・

DandDインスタンスによる解決



- Data and Description

データと，

その属性や補助情報など
データの解釈や説明も一緒に保存

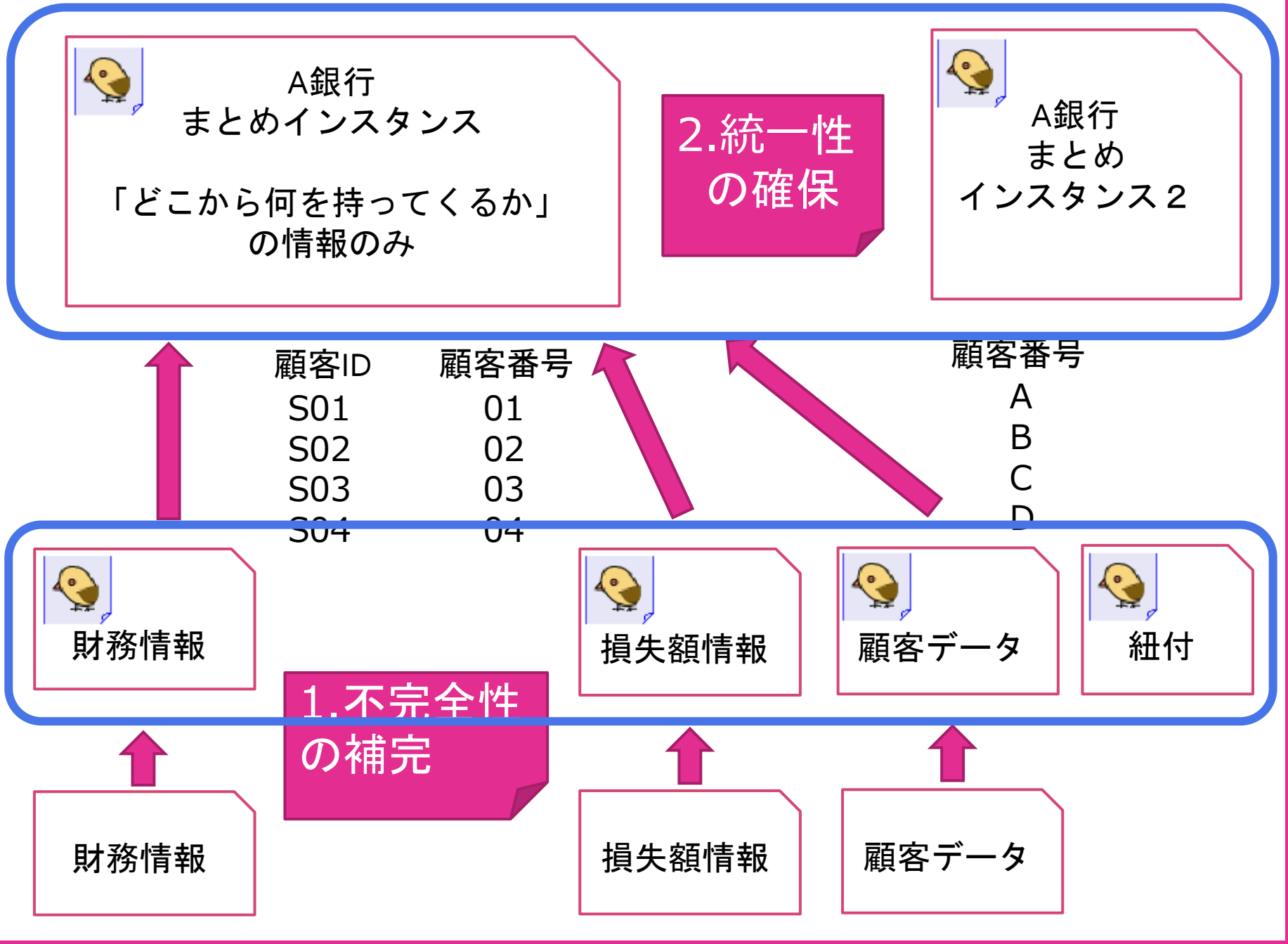
<Da
DataType="Measurement" Id="dv_1" LongName="dv
_1.In1" ShortNam
649 801628 81494
245828 618408 73
aVector>↓
<DataVector DataType="Measurement" Id="dv_2" LongName="dv
_2.In1" ShortName="v2" Unit="unit_1"> 20210331 20210331 2021
10331 20210331 20210331 20210331 20210331 20210331 2
20210331 20210331 20210331 20210331 20210331 2021033
1 20210331 20210331 20210331</DataVector>↓

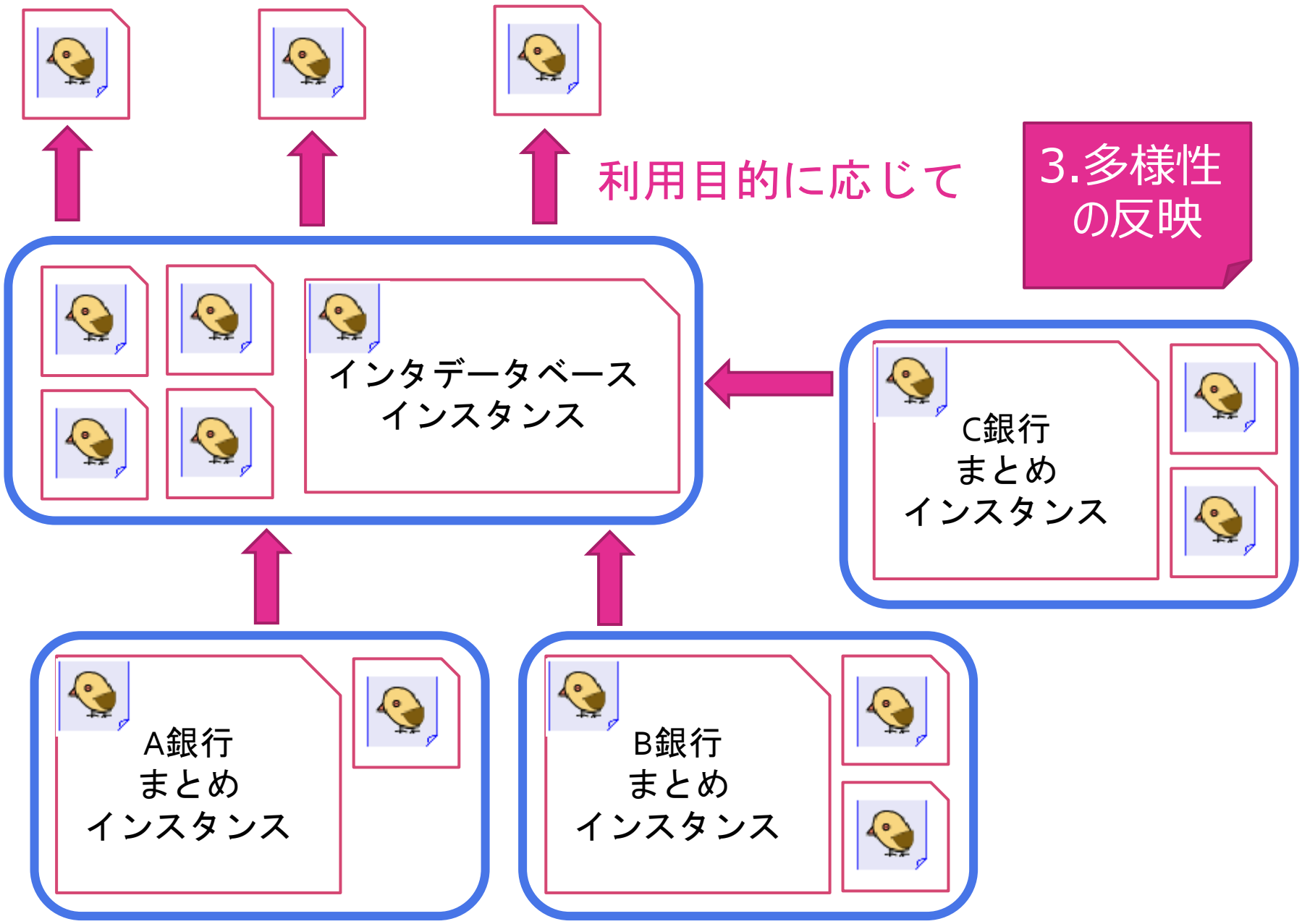
1次元

<DataVector DataType="Measurement" Id="dv_3" LongName="dv
_3.In1" ShortName="v3" Unit="unit_2"> 4 5 3 7 7 9 3 6 8 2 5
8 9 3 6 8 9 1 7</DataVector>↓

<DataVector DataType="Measurement" Id="dv_4" LongName="dv
_4.In1" ShortName="v4" Unit="unit_3"> 3 NA 4 NA NA NA NA NA
8 NA NA 8 9 NA 7 NA NA 1 7</DataVector>↓

- 処理は TRAD  を用いて行う.
  - XMLファイルなので汎用性がある.
- DandDインスタンスを通してデータを間接的に処理.
  - データの記述は一度だけで，データの解釈が書かれている.
- 複数のデータテーブルを保存可能.
  - 処理前後のデータテーブルの行き来が容易.
  - DandDインスタンスのサイズはほとんど変化しない.
- DandDルールがある.
  - 誰が編集しても同じ構造，読みやすい.


A銀行
まとめインスタンス
「どこから何を持ってくるか」
の情報のみ
2.統一性
の確保
A銀行
まとめ
インスタンス２
顧客ID
顧客番号
S01
01
S02
02
S03
03
S04
04
顧客番号
A
B
C
D
財務情報
損失額情報
顧客データ
紐付
1.不完全性
の補完
財務情報
損失額情報
顧客データ


利用目的に応じて
3.多様性
の反映
インタデータベース
インスタンス
C銀行
まとめ
インスタンス
A銀行
まとめ
インスタンス
B銀行
まとめ
インスタンス

これからのデータ活用

- 多ソース多サイトデータを活用するときの課題
  - 不完全性，不統一性，多様性
- DandDインスタンス群の作成
  - データとデータの解釈を共に記載
  - 動的な変化に柔軟に対応できる
  - 多種多様な目的にも対応できる
  - 透明性を確保できる
- 課題の根本的な解決を
  - 労力の無駄遣いをしない！