リレーショナルデータベースの高度利用環境DandD

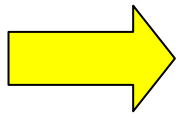
慶應義塾大学理工学部
横内大介
慶應義塾大学理工学部
柴田里程

データの高度利用

- データの公開
  - 情報公開
  - インターネットの普及，発展
  - データの配布形式，組織化の多様性
- 具体例
  - インターネットによる配布
    - LIBOR (London InterBank Offered Rate)
    - 地震データ
  - CD-ROMによる頒布
    - 気象データ（気象業務支援センター）
    - 民力（朝日新聞社）

データの高度利用を阻む障壁

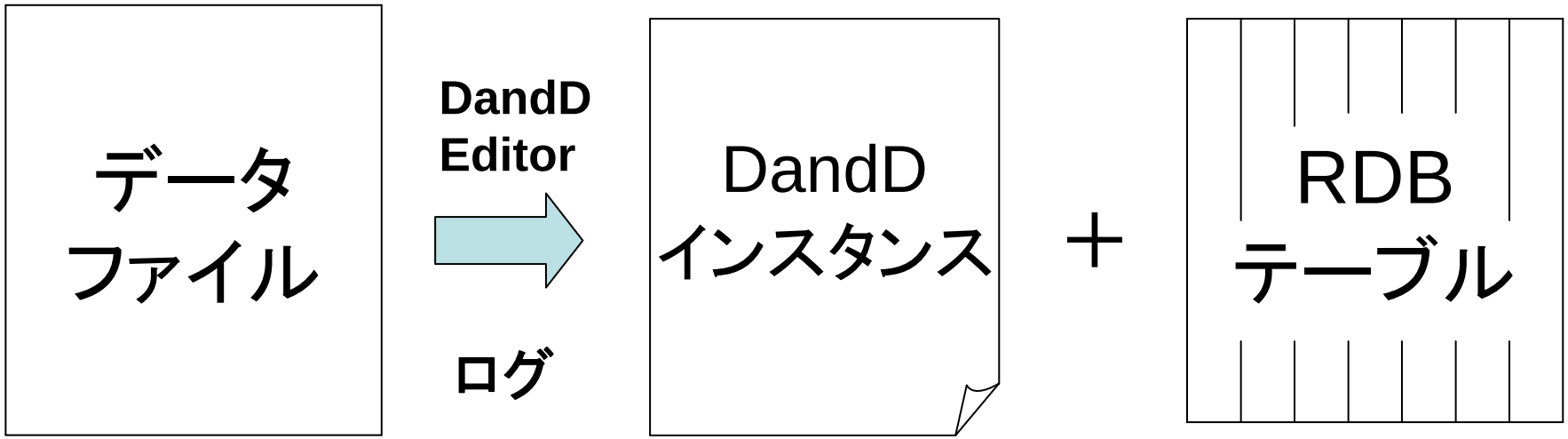
- 分割，分散
  - 公開側の都合
- 形式の多様性
  - 統一基準がない
- 規模
  - 情報技術の向上
- 信頼性
  - 保証する仕掛けがない
- 記述の曖昧さ
  - 客観的に評価するシステム
- 公開の原則の欠如
  - 風土



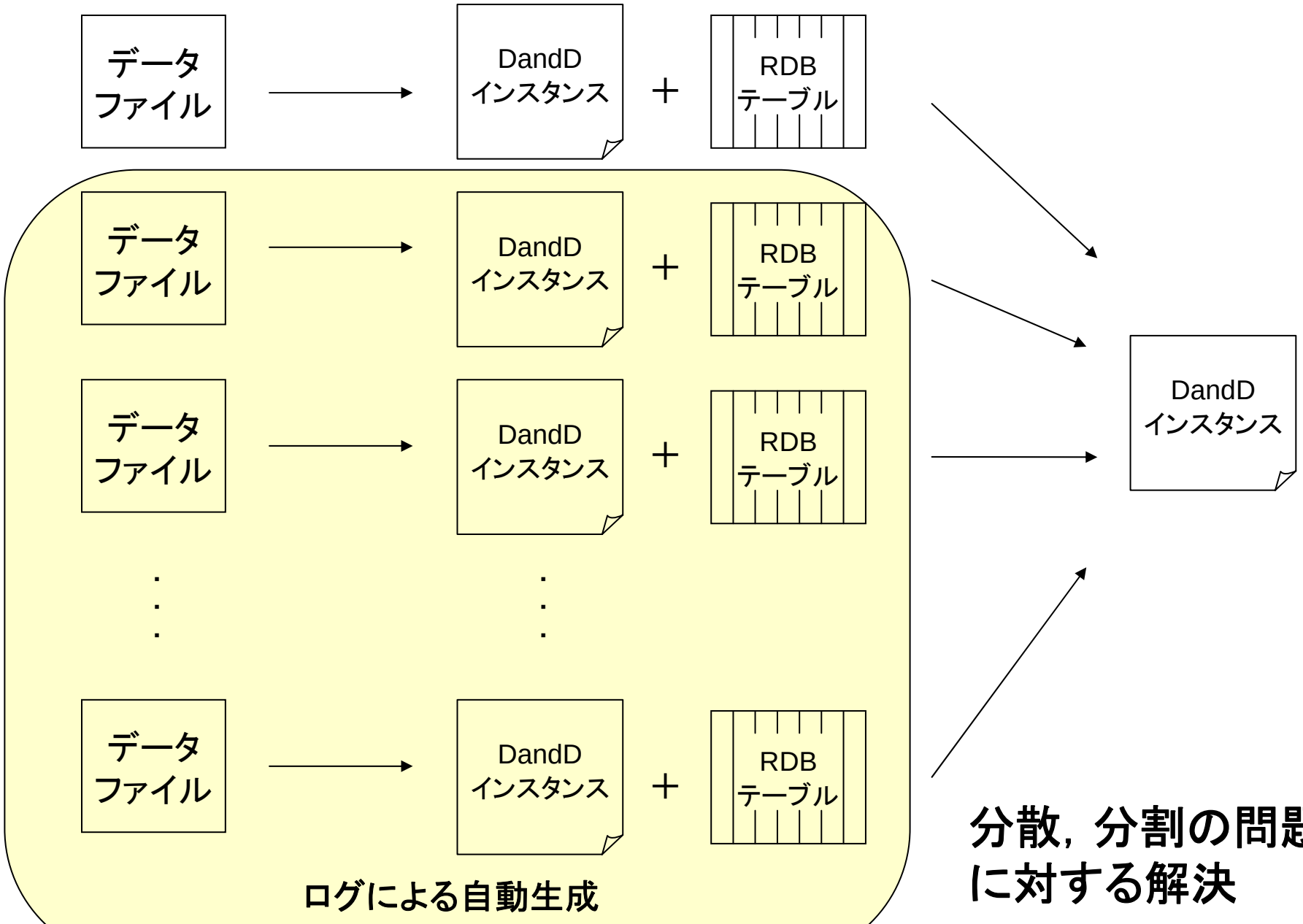
DandD環境による
5つの障壁の除去

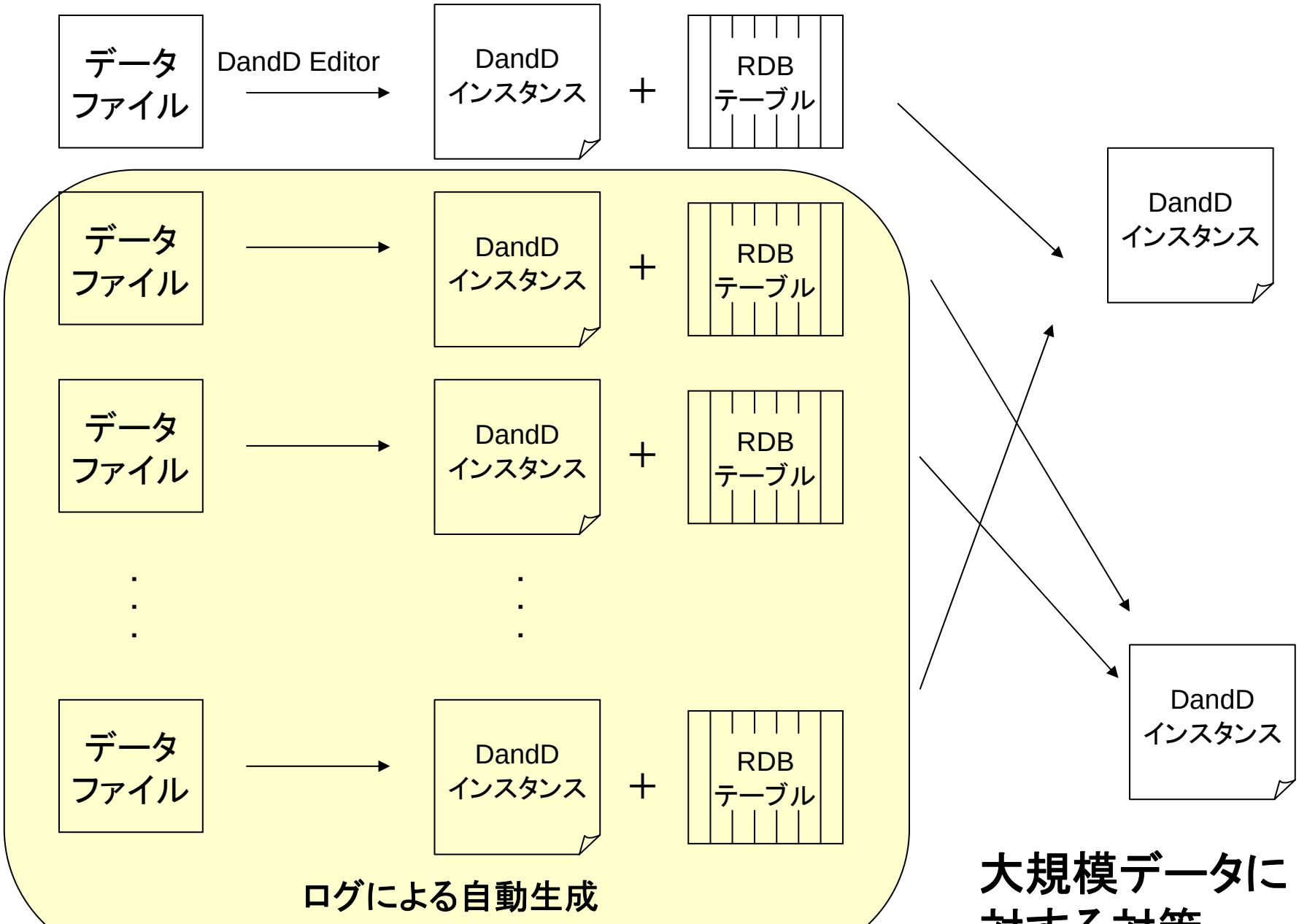
- DandDインスタンス
- リレーショナルデータベース

DandDインスタンスとRDB



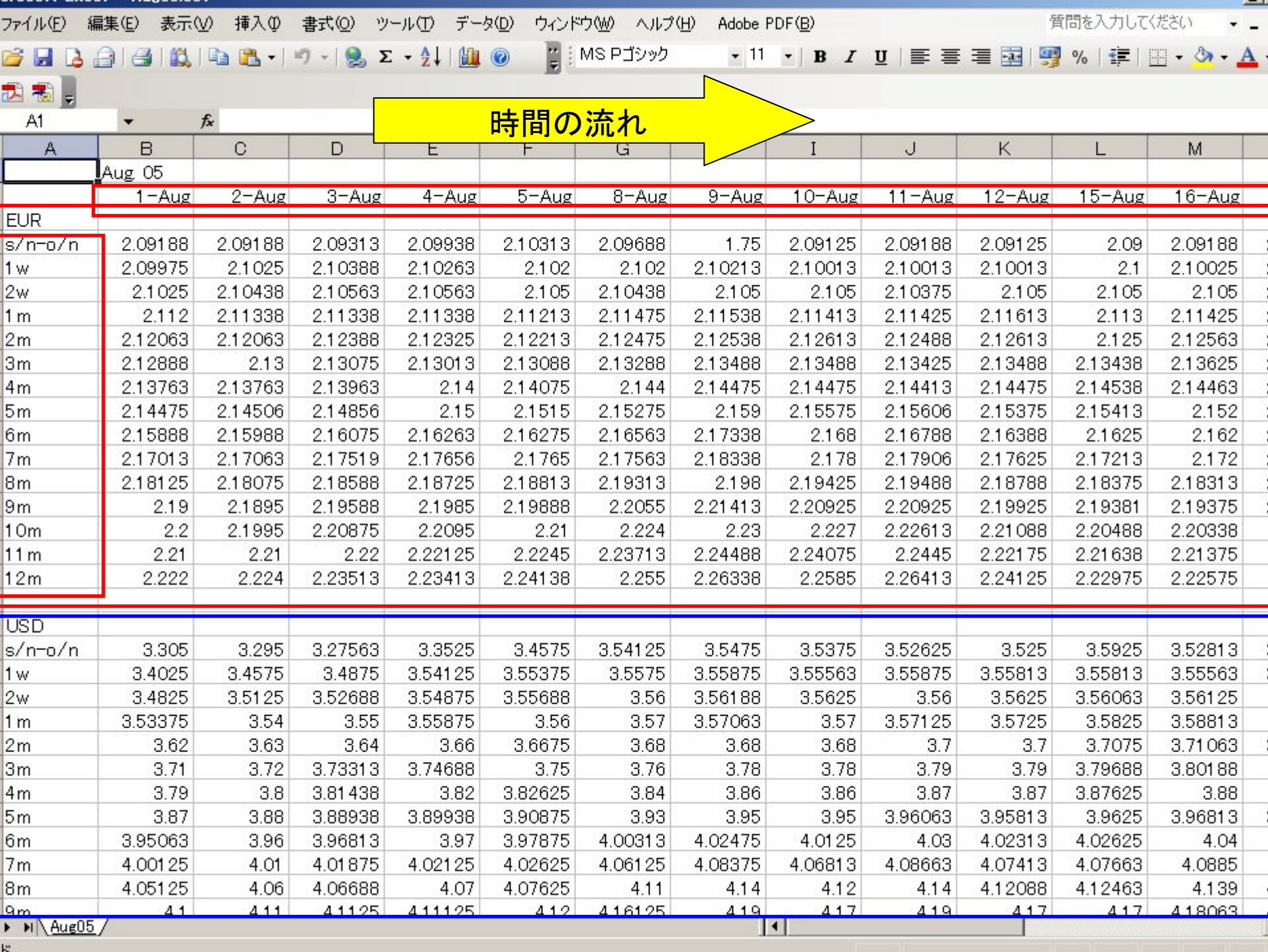
- 形式の多様性の吸収
- 信頼性の保証
- 形式的な記述


データ
ファイル
DandD
インスタンス
+
RDB
テーブル
データ
ファイル
DandD
インスタンス
+
RDB
テーブル
データ
ファイル
DandD
インスタンス
+
RDB
テーブル
データ
ファイル
DandD
インスタンス
+
RDB
テーブル
DandD
インスタンス
ログによる自動生成
分散，分割の問題
に対する解決


データ
ファイル
DandD Editor
DandD
インスタンス
＋
RDB
テーブル
データ
ファイル
DandD
インスタンス
＋
RDB
テーブル
データ
ファイル
DandD
インスタンス
＋
RDB
テーブル
データ
ファイル
DandD
インスタンス
＋
RDB
テーブル
DandD
インスタンス
DandD
インスタンス
ログによる自動生成
大規模データに
対する対策

実際例

- LIBOR
  - ロンドン市場での銀行間平均貸し手金利
  - BBA(The British Banker’s Association)が日に一度発表
  - 1ヶ月単位のExcelシート
  - シート 1 枚に9種類の金利データ
  - 構造が複雑
- モデルケース
  - 2005年1月から8月までのEURO金利を併合しR上で利用する



	A	B	C	D	E	F	G		I	J	K	L	M
		Aug 05											
		1-Aug	2-Aug	3-Aug	4-Aug	5-Aug	8-Aug	9-Aug	10-Aug	11-Aug	12-Aug	15-Aug	16-Aug
	EUR												
	s/n-o/n	2.09188	2.09188	2.09313	2.09938	2.10313	2.09688	1.75	2.09125	2.09188	2.09125	2.09	2.09188
	1w	2.09975	2.1025	2.10388	2.10263	2.102	2.102	2.10213	2.10013	2.10013	2.10013	2.1	2.10025
	2w	2.1025	2.10438	2.10563	2.10563	2.105	2.10438	2.105	2.105	2.10375	2.105	2.105	2.105
	1m	2.112	2.11338	2.11338	2.11338	2.11213	2.11475	2.11538	2.11413	2.11425	2.11613	2.113	2.11425
	2m	2.12063	2.12063	2.12388	2.12325	2.12213	2.12475	2.12538	2.12613	2.12488	2.12613	2.125	2.12563
	3m	2.12888	2.13	2.13075	2.13013	2.13088	2.13288	2.13488	2.13488	2.13425	2.13488	2.13438	2.13625
	4m	2.13763	2.13763	2.13963	2.14	2.14075	2.144	2.14475	2.14475	2.14413	2.14475	2.14538	2.14463
	5m	2.14475	2.14506	2.14856	2.15	2.1515	2.15275	2.159	2.15575	2.15606	2.15375	2.15413	2.152
	6m	2.15888	2.15988	2.16075	2.16263	2.16275	2.16563	2.17338	2.168	2.16788	2.16388	2.1625	2.162
	7m	2.17013	2.17063	2.17519	2.17656	2.1765	2.17563	2.18338	2.178	2.17906	2.17625	2.17213	2.172
	8m	2.18125	2.18075	2.18588	2.18725	2.18813	2.19313	2.198	2.19425	2.19488	2.18788	2.18375	2.18313
	9m	2.19	2.1895	2.19588	2.1985	2.19888	2.2055	2.21413	2.20925	2.20925	2.19925	2.19381	2.19375
	10m	2.2	2.1995	2.20875	2.2095	2.21	2.224	2.23	2.227	2.22613	2.21088	2.20488	2.20338
	11m	2.21	2.21	2.22	2.22125	2.2245	2.23713	2.24488	2.24075	2.2445	2.22175	2.21638	2.21375
	12m	2.222	2.224	2.23513	2.23413	2.24138	2.255	2.26338	2.2585	2.26413	2.24125	2.22975	2.22575
	USD												
	s/n-o/n	3.305	3.295	3.27563	3.3525	3.4575	3.54125	3.5475	3.5375	3.52625	3.525	3.5925	3.52813
	1w	3.4025	3.4575	3.4875	3.54125	3.55375	3.5575	3.55875	3.55563	3.55875	3.55813	3.55813	3.55563
	2w	3.4825	3.5125	3.52688	3.54875	3.55688	3.56	3.56188	3.5625	3.56	3.5625	3.56063	3.56125
	1m	3.53375	3.54	3.55	3.55875	3.56	3.57	3.57063	3.57	3.57125	3.5725	3.5825	3.58813
	2m	3.62	3.63	3.64	3.66	3.6675	3.68	3.68	3.68	3.7	3.7	3.7075	3.71063
	3m	3.71	3.72	3.73313	3.74688	3.75	3.76	3.78	3.78	3.79	3.79	3.79688	3.80188
	4m	3.79	3.8	3.81438	3.82	3.82625	3.84	3.86	3.86	3.87	3.87	3.87625	3.88
	5m	3.87	3.88	3.88938	3.89938	3.90875	3.93	3.95	3.95	3.96063	3.95813	3.9625	3.96813
	6m	3.95063	3.96	3.96813	3.97	3.97875	4.00313	4.02475	4.0125	4.03	4.02313	4.02625	4.04
	7m	4.00125	4.01	4.01875	4.02125	4.02625	4.06125	4.08375	4.06813	4.08663	4.07413	4.07663	4.0885
	8m	4.05125	4.06	4.06688	4.07	4.07625	4.11	4.14	4.12	4.14	4.12088	4.12463	4.139
	9m	4.1	4.11	4.11125	4.11125	4.12	4.16125	4.19	4.17	4.19	4.17	4.17	4.18063

表計算ソフトウェアとRを用いた一般的な作業工程

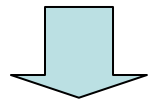
表計算ソフト

- データの切り出し
- データの転置
- 複数のシートを結合
- 適切なヘッダーの付与
- R への取り込み

DandD環境

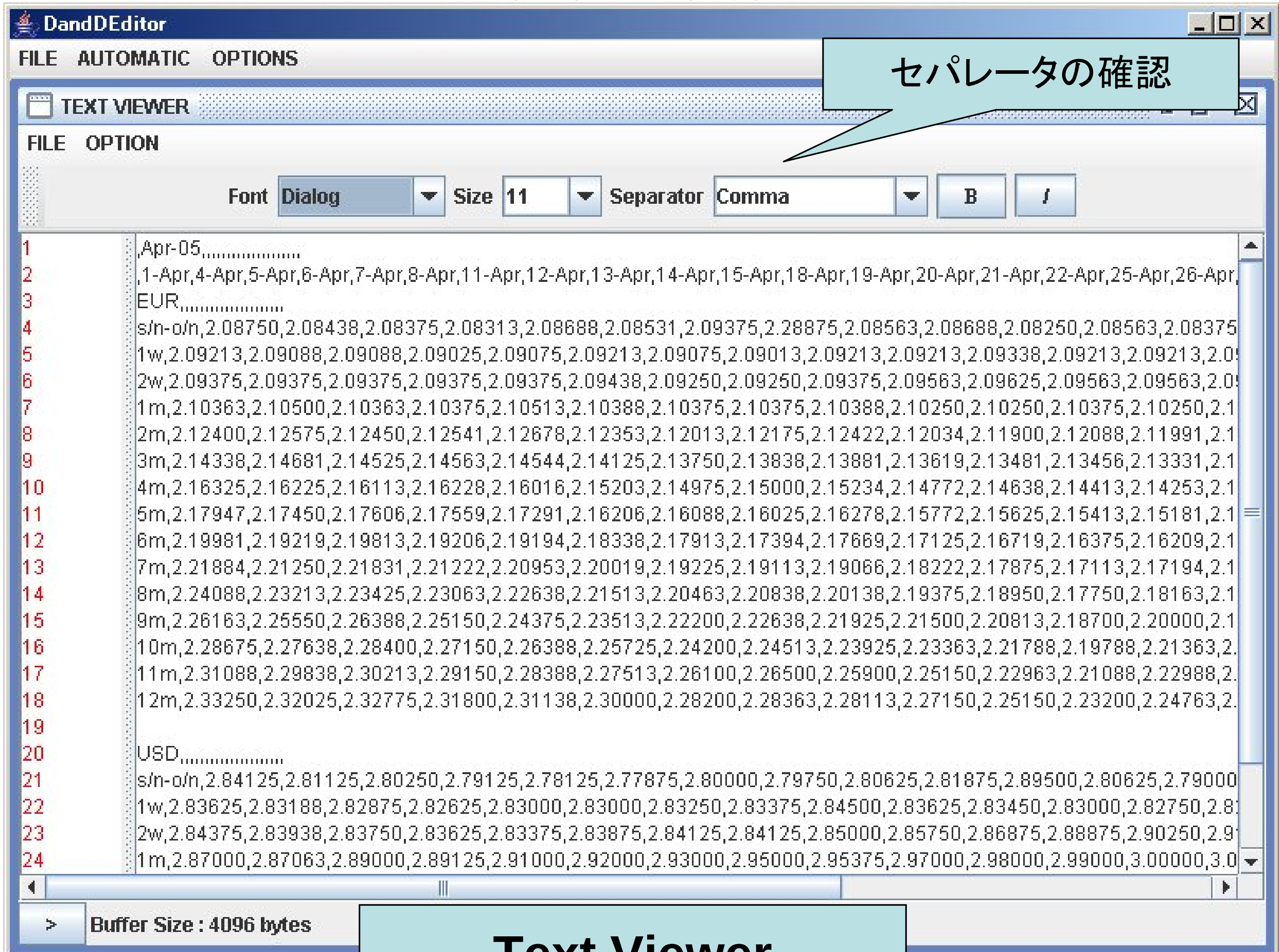
DandD

- データの切り出し
- データの転置
- DandDインスタンスの生成
- R への取り込み



DandDEditor と DandDR

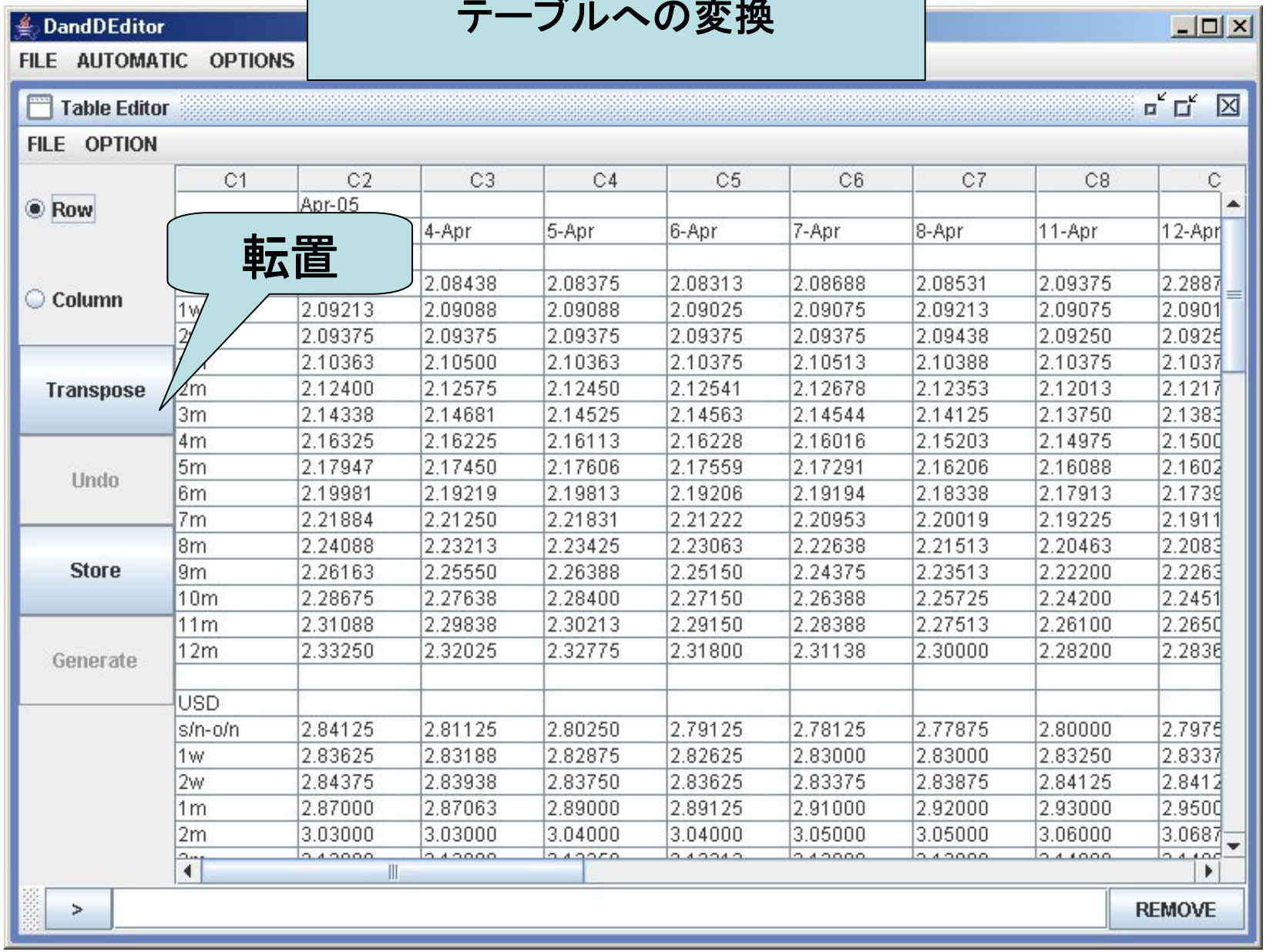
DandD Editor

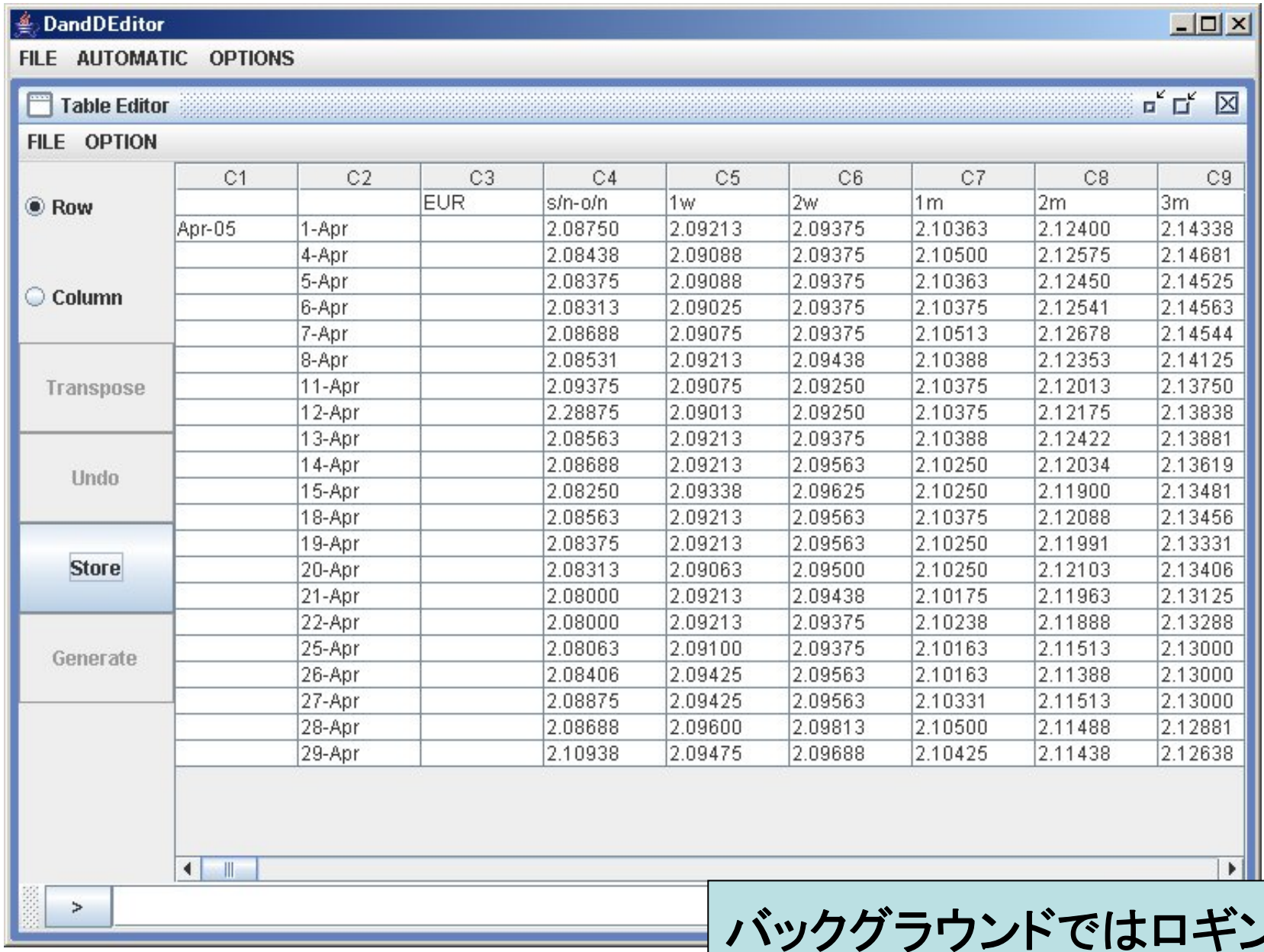


セパレータの確認

Text Viewer

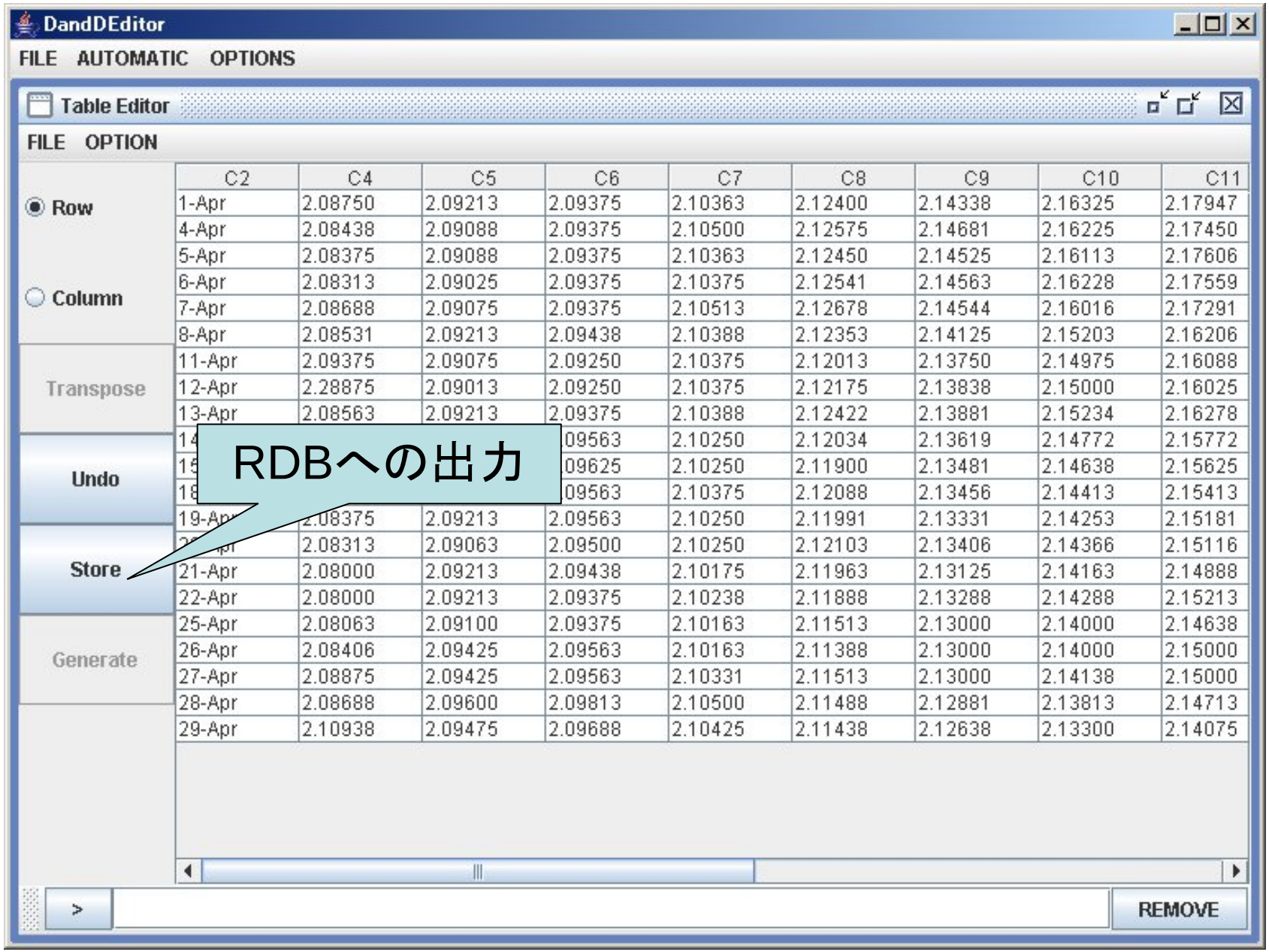
テーブルへの変換

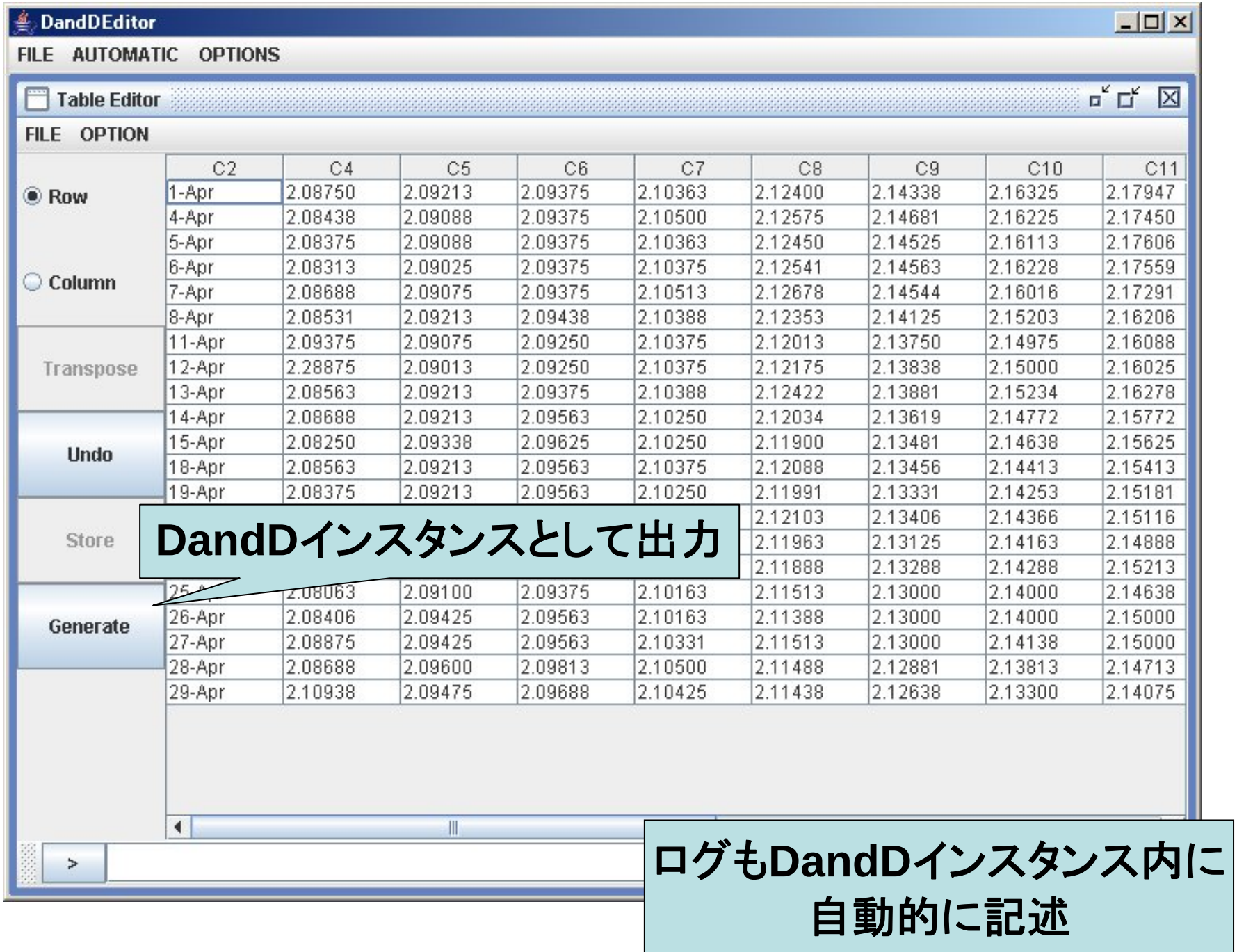

DandDEditor
FILE AUTOMATIC OPTIONS
Table Editor
FILE OPTION
Row
Column
Transpose
Undo
Store
Generate
転置
C1 C2 C3 C4 C5 C6 C7 C8 C
Apr-05
4-Apr 5-Apr 6-Apr 7-Apr 8-Apr 11-Apr 12-Apr
2.08438 2.08375 2.08313 2.08688 2.08531 2.09375 2.2887
1w 2.09213 2.09088 2.09088 2.09025 2.09075 2.09213 2.09075 2.0901
2.09375 2.09375 2.09375 2.09375 2.09375 2.09438 2.09250 2.0925
2.10363 2.10500 2.10363 2.10375 2.10513 2.10388 2.10375 2.1037
2m 2.12400 2.12575 2.12450 2.12541 2.12678 2.12353 2.12013 2.1217
3m 2.14338 2.14681 2.14525 2.14563 2.14544 2.14125 2.13750 2.1383
4m 2.16325 2.16225 2.16113 2.16228 2.16016 2.15203 2.14975 2.1500
5m 2.17947 2.17450 2.17606 2.17559 2.17291 2.16206 2.16088 2.1602
6m 2.19981 2.19219 2.19813 2.19206 2.19194 2.18338 2.17913 2.1739
7m 2.21884 2.21250 2.21831 2.21222 2.20953 2.20019 2.19225 2.1911
8m 2.24088 2.23213 2.23425 2.23063 2.22638 2.21513 2.20463 2.2083
9m 2.26163 2.25550 2.26388 2.25150 2.24375 2.23513 2.22200 2.2263
10m 2.28675 2.27638 2.28400 2.27150 2.26388 2.25725 2.24200 2.2451
11m 2.31088 2.29838 2.30213 2.29150 2.28388 2.27513 2.26100 2.2650
12m 2.33250 2.32025 2.32775 2.31800 2.31138 2.30000 2.28200 2.2836
USD
s/n-o/n 2.84125 2.81125 2.80250 2.79125 2.78125 2.77875 2.80000 2.7975
1w 2.83625 2.83188 2.82875 2.82625 2.83000 2.83000 2.83250 2.8337
2w 2.84375 2.83938 2.83750 2.83625 2.83375 2.83875 2.84125 2.8412
1m 2.87000 2.87063 2.89000 2.89125 2.91000 2.92000 2.93000 2.9500
2m 3.03000 3.03000 3.04000 3.04000 3.05000 3.05000 3.06000 3.0687
>
REMOVE

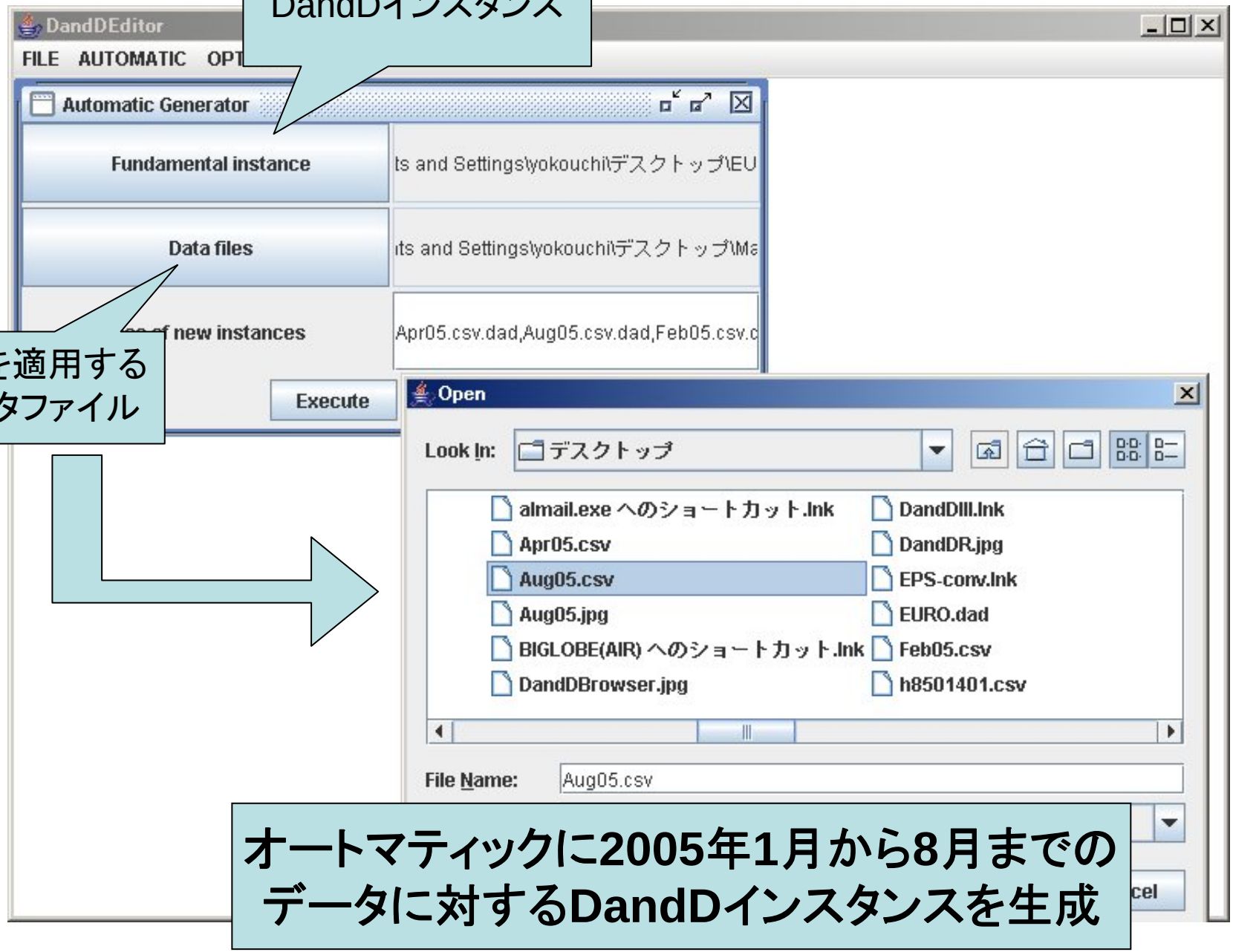


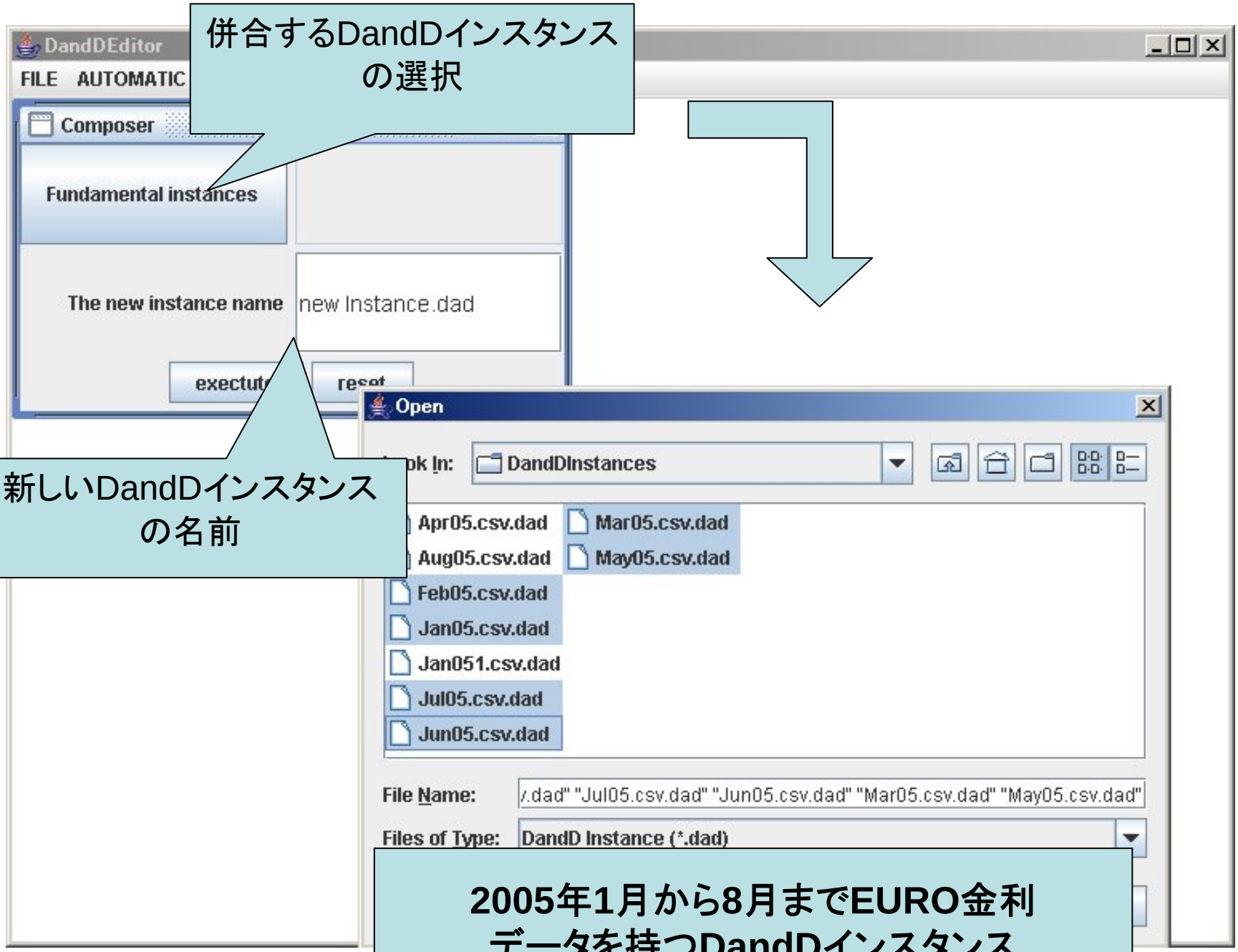
C1	C2	C3	C4	C5	C6	C7	C8	C9
		EUR	s/n-o/n	1w	2w	1m	2m	3m
Apr-05	1-Apr		2.08750	2.09213	2.09375	2.10363	2.12400	2.14338
	4-Apr		2.08438	2.09088	2.09375	2.10500	2.12575	2.14681
	5-Apr		2.08375	2.09088	2.09375	2.10363	2.12450	2.14525
	6-Apr		2.08313	2.09025	2.09375	2.10375	2.12541	2.14563
	7-Apr		2.08688	2.09075	2.09375	2.10513	2.12678	2.14544
	8-Apr		2.08531	2.09213	2.09438	2.10388	2.12353	2.14125
	11-Apr		2.09375	2.09075	2.09250	2.10375	2.12013	2.13750
	12-Apr		2.28875	2.09013	2.09250	2.10375	2.12175	2.13838
	13-Apr		2.08563	2.09213	2.09375	2.10388	2.12422	2.13881
	14-Apr		2.08688	2.09213	2.09563	2.10250	2.12034	2.13619
	15-Apr		2.08250	2.09338	2.09625	2.10250	2.11900	2.13481
	18-Apr		2.08563	2.09213	2.09563	2.10375	2.12088	2.13456
	19-Apr		2.08375	2.09213	2.09563	2.10250	2.11991	2.13331
	20-Apr		2.08313	2.09063	2.09500	2.10250	2.12103	2.13406
	21-Apr		2.08000	2.09213	2.09438	2.10175	2.11963	2.13125
	22-Apr		2.08000	2.09213	2.09375	2.10238	2.11888	2.13288
	25-Apr		2.08063	2.09100	2.09375	2.10163	2.11513	2.13000
	26-Apr		2.08406	2.09425	2.09563	2.10163	2.11388	2.13000
	27-Apr		2.08875	2.09425	2.09563	2.10331	2.11513	2.13000
	28-Apr		2.08688	2.09600	2.09813	2.10500	2.11488	2.12881
	29-Apr		2.10938	2.09475	2.09688	2.10425	2.11438	2.12638


DandDEditor
FILE AUTOMATIC OPTIONS
不要な列や行の削除
Table Editor
FILE OPTION
Row
Column
Transpose
Undo
Store
Generate
C1 C2 C3 C4 C5 C6 C7 C8 C9
EUR s/n-o/n 1w 2w 1m 2m 3m
Apr-05 1-Apr 2.08750 2.09213 2.09375 2.10363 2.12400 2.14338
4-Apr 2.08438 2.09088 2.09375 2.10500 2.12575 2.14681
5-Apr 2.08375 2.09088 2.09375 2.10363 2.12450 2.14525
6-Apr 2.08313 2.09025 2.09375 2.10375 2.12541 2.14563
7-Apr 2.08688 2.09075 2.09375 2.10513 2.12678 2.14544
8-Apr 2.08531 2.09213 2.09438 2.10388 2.12353 2.14125
11-Apr 2.09375 2.09075 2.09250 2.10375 2.12013 2.13750
12-Apr 2.28875 2.09013 2.09250 2.10375 2.12175 2.13838
13-Apr 2.08563 2.09213 2.09375 2.10388 2.12422 2.13881
14-Apr 2.08688 2.09213 2.09563 2.10250 2.12034 2.13619
15-Apr 2.08250 2.09338 2.09625 2.10250 2.11900 2.13481
18-Apr 2.08563 2.09213 2.09563 2.10375 2.12088 2.13456
19-Apr 2.08375 2.09213 2.09563 2.10250 2.11991 2.13331
20-Apr 2.08313 2.09063 2.09500 2.10250 2.12103 2.13406
21-Apr 2.08000 2.09213 2.09438 2.10175 2.11963 2.13125
22-Apr 2.08000 2.09213 2.09375 2.10238 2.11888 2.13288
25-Apr 2.08063 2.09100 2.09375 2.10163 2.11513 2.13000
26-Apr 2.08406 2.09425 2.09563 2.10163 2.11388 2.13000
27-Apr 2.08875 2.09425 2.09563 2.10331 2.11513 2.13000
28-Apr 2.08688 2.09600 2.09813 2.10500 2.11488 2.12881
29-Apr 2.10938 2.09475 2.09688 2.10425 2.11438 2.12638
>
バックグラウンドではロギング

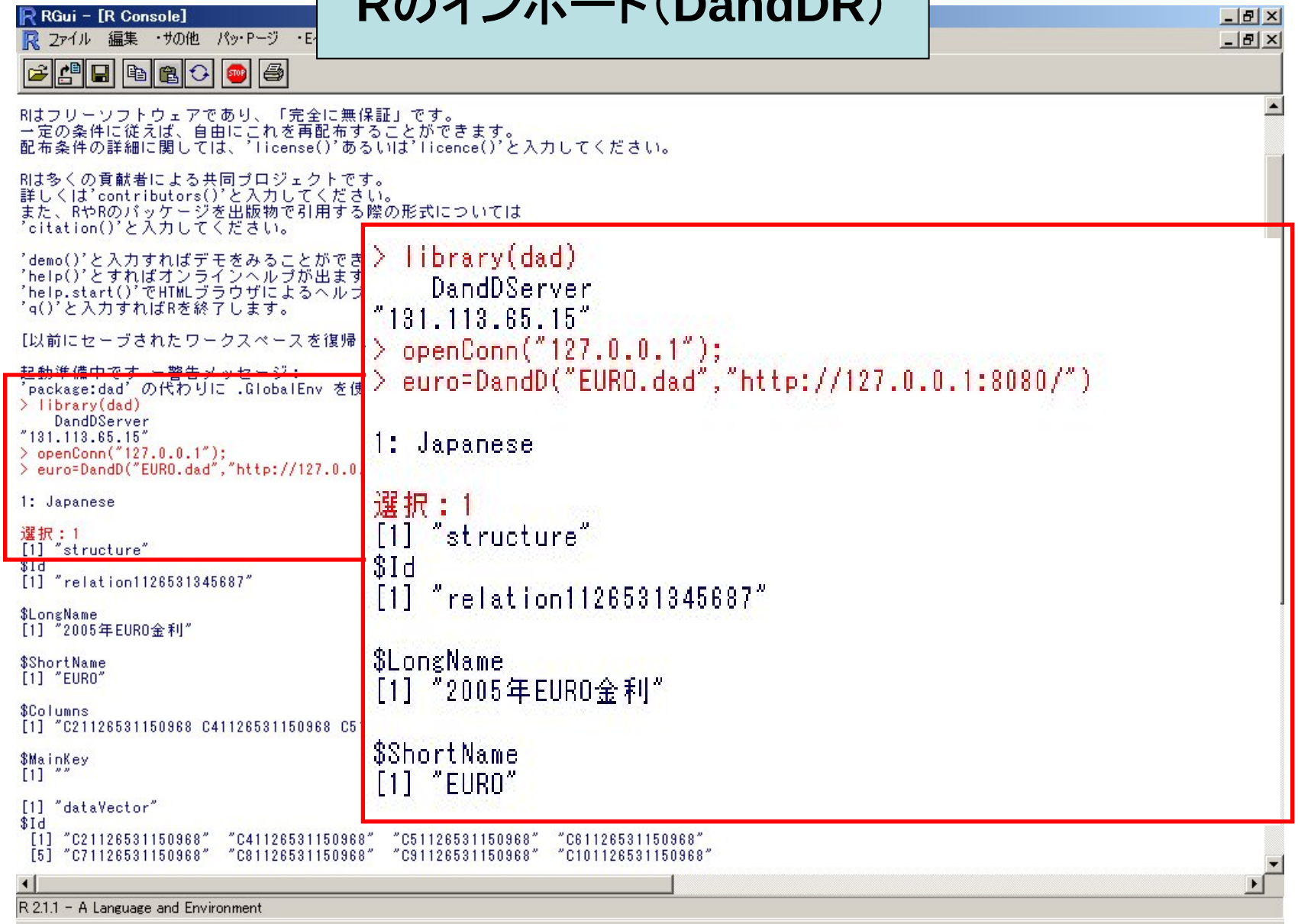

DandDEditor
FILE AUTOMATIC OPTIONS
Table Editor
FILE OPTION
Row
Column
Transpose
Undo
Store
Generate
RDBへの出力
C2 C4 C5 C6 C7 C8 C9 C10 C11
1-Apr 2.08750 2.09213 2.09375 2.10363 2.12400 2.14338 2.16325 2.17947
4-Apr 2.08438 2.09088 2.09375 2.10500 2.12575 2.14681 2.16225 2.17450
5-Apr 2.08375 2.09088 2.09375 2.10363 2.12450 2.14525 2.16113 2.17606
6-Apr 2.08313 2.09025 2.09375 2.10375 2.12541 2.14563 2.16228 2.17559
7-Apr 2.08688 2.09075 2.09375 2.10513 2.12678 2.14544 2.16016 2.17291
8-Apr 2.08531 2.09213 2.09438 2.10388 2.12353 2.14125 2.15203 2.16206
11-Apr 2.09375 2.09075 2.09250 2.10375 2.12013 2.13750 2.14975 2.16088
12-Apr 2.28875 2.09013 2.09250 2.10375 2.12175 2.13838 2.15000 2.16025
13-Apr 2.08563 2.09213 2.09375 2.10388 2.12422 2.13881 2.15234 2.16278
2.10250 2.12034 2.13619 2.14772 2.15772
2.10250 2.11900 2.13481 2.14638 2.15625
2.10375 2.12088 2.13456 2.14413 2.15413
19-Apr 2.08375 2.09213 2.09563 2.10250 2.11991 2.13331 2.14253 2.15181
2.08313 2.09063 2.09500 2.10250 2.12103 2.13406 2.14366 2.15116
21-Apr 2.08000 2.09213 2.09438 2.10175 2.11963 2.13125 2.14163 2.14888
22-Apr 2.08000 2.09213 2.09375 2.10238 2.11888 2.13288 2.14288 2.15213
25-Apr 2.08063 2.09100 2.09375 2.10163 2.11513 2.13000 2.14000 2.14638
26-Apr 2.08406 2.09425 2.09563 2.10163 2.11388 2.13000 2.14000 2.15000
27-Apr 2.08875 2.09425 2.09563 2.10331 2.11513 2.13000 2.14138 2.15000
28-Apr 2.08688 2.09600 2.09813 2.10500 2.11488 2.12881 2.13813 2.14713
29-Apr 2.10938 2.09475 2.09688 2.10425 2.11438 2.12638 2.13300 2.14075
>
REMOVE


DandDEditor
FILE AUTOMATIC OPTIONS
Table Editor
FILE OPTION
Row
Column
Transpose
Undo
Store
Generate
C2 C4 C5 C6 C7 C8 C9 C10 C11
1-Apr 2.08750 2.09213 2.09375 2.10363 2.12400 2.14338 2.16325 2.17947
4-Apr 2.08438 2.09088 2.09375 2.10500 2.12575 2.14681 2.16225 2.17450
5-Apr 2.08375 2.09088 2.09375 2.10363 2.12450 2.14525 2.16113 2.17606
6-Apr 2.08313 2.09025 2.09375 2.10375 2.12541 2.14563 2.16228 2.17559
7-Apr 2.08688 2.09075 2.09375 2.10513 2.12678 2.14544 2.16016 2.17291
8-Apr 2.08531 2.09213 2.09438 2.10388 2.12353 2.14125 2.15203 2.16206
11-Apr 2.09375 2.09075 2.09250 2.10375 2.12013 2.13750 2.14975 2.16088
12-Apr 2.28875 2.09013 2.09250 2.10375 2.12175 2.13838 2.15000 2.16025
13-Apr 2.08563 2.09213 2.09375 2.10388 2.12422 2.13881 2.15234 2.16278
14-Apr 2.08688 2.09213 2.09563 2.10250 2.12034 2.13619 2.14772 2.15772
15-Apr 2.08250 2.09338 2.09625 2.10250 2.11900 2.13481 2.14638 2.15625
18-Apr 2.08563 2.09213 2.09563 2.10375 2.12088 2.13456 2.14413 2.15413
19-Apr 2.08375 2.09213 2.09563 2.10250 2.11991 2.13331 2.14253 2.15181
2.12103 2.13406 2.14366 2.15116
2.11963 2.13125 2.14163 2.14888
2.11888 2.13288 2.14288 2.15213
2.08063 2.09100 2.09375 2.10163 2.11513 2.13000 2.14000 2.14638
26-Apr 2.08406 2.09425 2.09563 2.10163 2.11388 2.13000 2.14000 2.15000
27-Apr 2.08875 2.09425 2.09563 2.10331 2.11513 2.13000 2.14138 2.15000
28-Apr 2.08688 2.09600 2.09813 2.10500 2.11488 2.12881 2.13813 2.14713
29-Apr 2.10938 2.09475 2.09688 2.10425 2.11438 2.12638 2.13300 2.14075
>
DandDインスタンスとして出力
ログもDandDインスタンス内に
自動的に記述


ログが書かれた
DandDインスタンス
DandDEditor
FILE AUTOMATIC OPT
Automatic Generator
Fundamental instance
ts and Settings\yokouchi\デスクトップ\EU
Data files
ts and Settings\yokouchi\デスクトップ\Ma
new instances
Apr05.csv.dad,Aug05.csv.dad,Feb05.csv.c
Execute
ログを適用する
データファイル
Open
Look In:
デスクトップ
almail.exe へのショートカット.lnk
Apr05.csv
Aug05.csv
Aug05.jpg
BIGLOBE(AIR) へのショートカット.lnk
DandDBrowser.jpg
DandDIII.lnk
DandDR.jpg
EPS-conv.lnk
EURO.dad
Feb05.csv
h8501401.csv
File Name:
Aug05.csv
オートマティックに2005年1月から8月までの
データに対するDandDインスタンスを生成

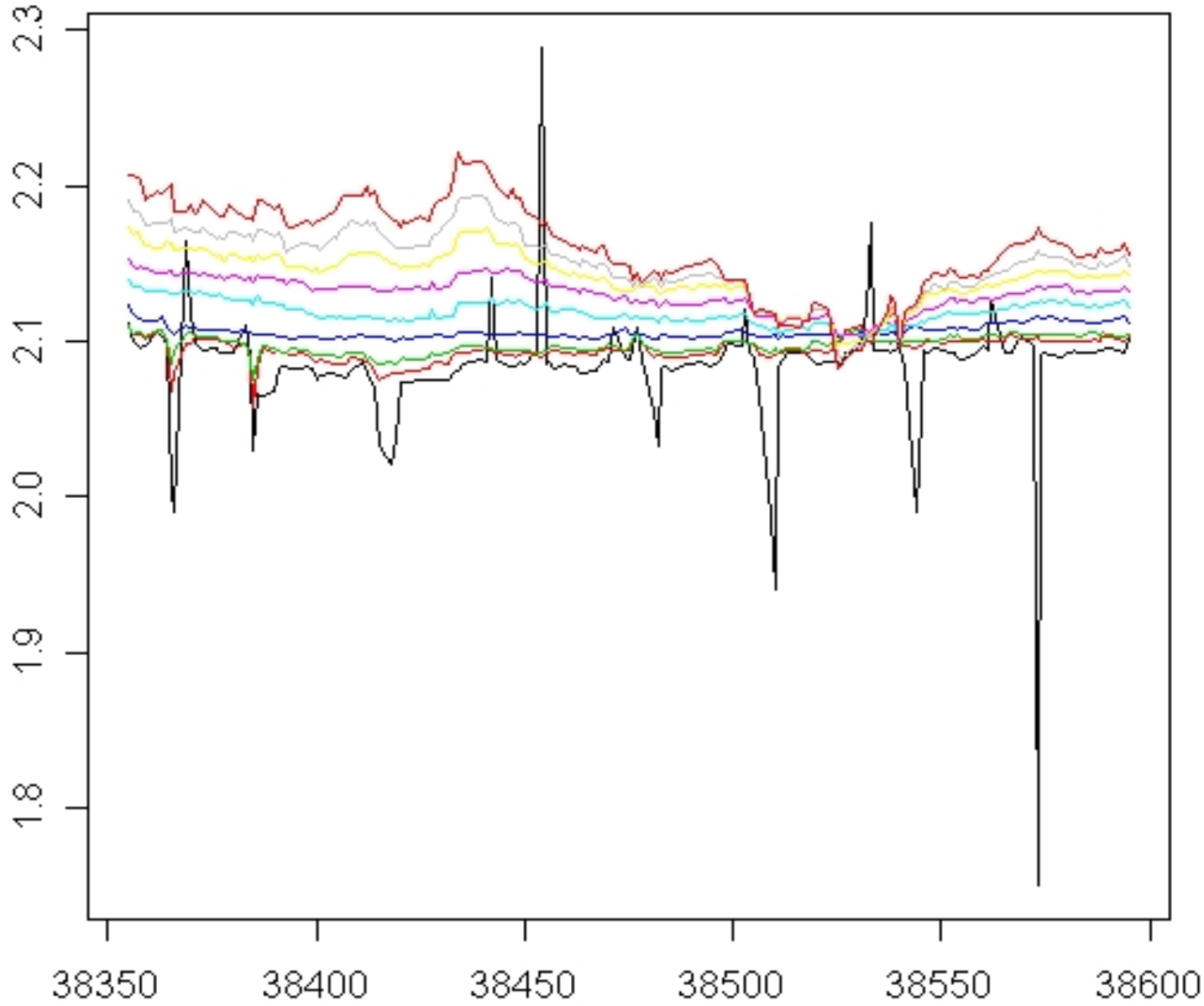

併合するDandDインスタンスの選択
DandDEditor
FILE AUTOMATIC
Composer
Fundamental instances
The new instance name
new Instance.dad
Open
DandDInstances
Apr05.csv.dad
Aug05.csv.dad
Feb05.csv.dad
Jan05.csv.dad
Jan051.csv.dad
Jul05.csv.dad
Jun05.csv.dad
Mar05.csv.dad
May05.csv.dad
File Name:
Files of Type:
DandD Instance (*.dad)
新しいDandDインスタンスの名前
2005年1月から8月までEURO金利データを持つDandDインスタンス

Rのインポート（DandDR）




```
> library(dad)
    DandDServer
"131.113.65.15"
> openConn("127.0.0.1");
> euro=DandD("EURO.dad","http://127.0.0.1:8080/")

1: Japanese

選択：1
[1] "structure"
$Id
[1] "relation1126531345687"

$LongName
[1] "2005年EURO金利"

$ShortName
[1] "EURO"
```

RによるEURO金利の時系列図



表計算ソフトウェアとRを用いた一般的な作業工程

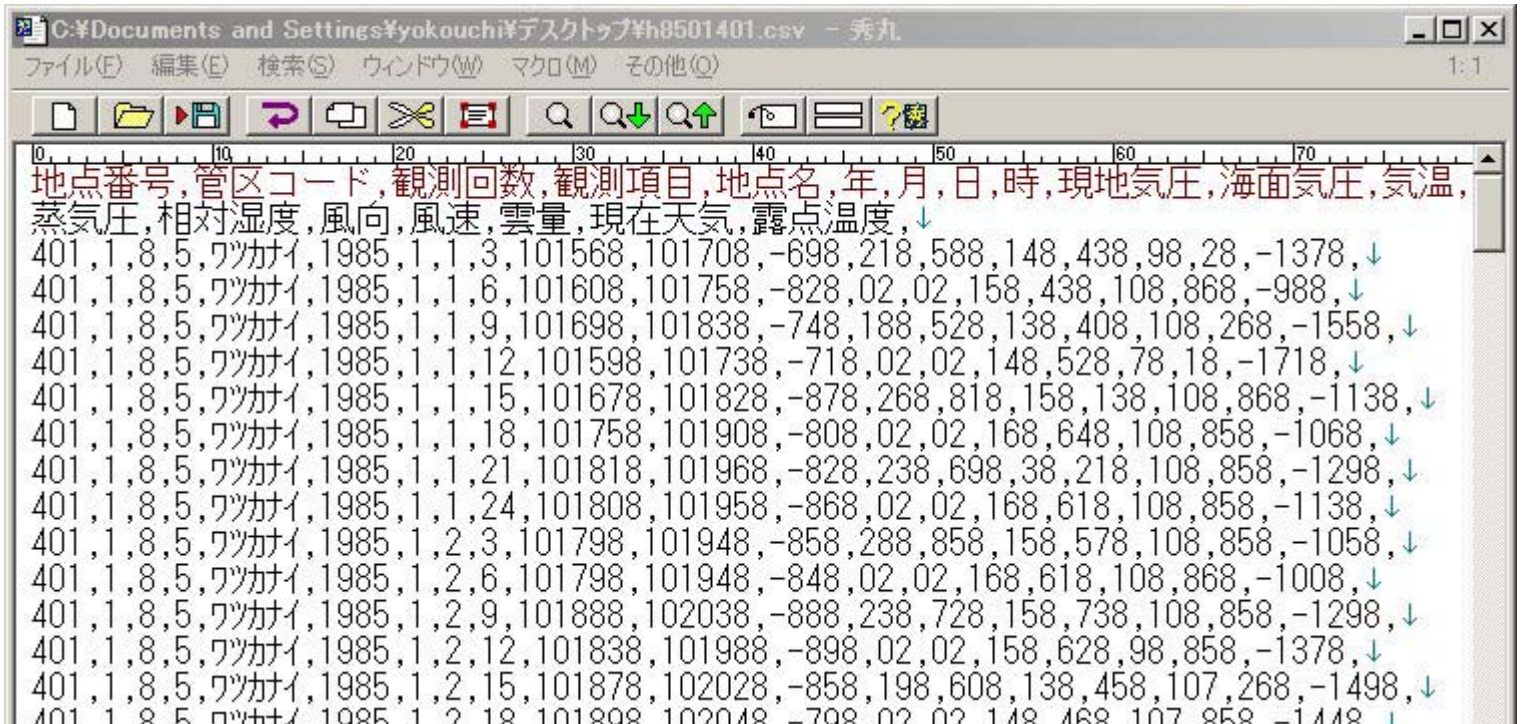
- 手作業中心
  - 表計算ソフトウェアやRの十分な知識
- 編集作業者依存
  - 妥当性
  - インフォーマルな編集記録

DandD環境

- DandDインスタンス化中心
  - DandDインスタンスにする手続きのみに特化
- オートマチックなログ機能
  - 信頼性の確保
  - 編集手続きの自動化

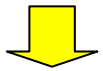
実際例2

- 気象データの利用
  - 1961年から2003年まで
  - 144観測所の月別データ
  - 観測間隔 3 時間ごと
  - 約61000ファイル，約 1.16 GB



表計算ソフトウェアとRを用いた一般的な作業工程

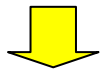
- データの編集
- 適切なヘッダー
- 複数のシートを結合
- R への取り込み



- 大規模データに対する限界
  - 表計算ソフト
    - 一部だけ
  - デフォルト設定のR
    - 読み込めるがおよそ1年分のデータで動作が不安定

DandD環境

- データの編集
- DandDインスタンスの生成
- R への取り込み（DandDR）



- 大規模データへの対応
  - 従来のDandD
    - RDB を背景に大規模データのストレージには対応
    - 無理やり1つのDandDインスタンスにまとめていた
  - 現在のDandD
    - 複数のDandDインスタンス
    - 必要な段階でまとめる

リレーショナルデータベース＋DandD

- データの高度利用
  - 表計算ソフトウェアよ，さようなら
    - 報告書に載せる表の作成が基本
      - モデルを創ることまでは考えていない
    - 手作業
    - 大規模データは扱えない
    - 編集手続き
  - DandDよ，こんにちは
    - 大規模データの扱い
      - 表計算ソフトやRだけでは限界
      - データのストレージはすべてRDB
    - Audit（監査）
      - フォーマル
      - 容易な編集の実現

DandDプロジェクトのホームページ

http://www.stat.math.keio.ac.jp/DandD/

必須属性（人間の手を煩わすか否か）

- DataVector（最小単位）の必須属性
  - データベクトル 1 本の外形を知る上で最低限必要なもの
  - 将来にわたって変更する必要性がないもの（言語の追加を除いて）
  - 個々の値に関わらない属性
- 構造(Relation) の必須属性
  - 1 つのRelationを外形を知る上で最低限必要なもの
  - 将来にわたって変更する必要性がないもの（言語の追加を除いて）
  - 個々のカラムの役割や結びつきに関わらない属性
- インスタンスの必須属性
  - インスタンス（データ全体）の外形を知る上で最低限必要なもの
  - 将来にわたって変更する必要性がないもの（言語の追加を除いて）
  - 個々のデータ構造に関わらない属性