官　活

（株）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

1　官庁データ活用の壁

我が国でも，ようやく e-Stat から諸官庁の公表データを一元的にデジタルデータとしてダウンロードできるようになってきたが，その本格的な活用となるとそう簡単ではない．基本的にこれまで印刷物として刊行してきた形式を踏襲しているため，フィジカルなデータ形式だけでなく，表現形式もさまざまである．一方で，RESAS のように目的別に GUI を構築してデータの活用を促す政府の動きもあるが，本格的な活用へのひとつの入口としては有効ではあるものの，そこから少しでも深堀りしようと思えば，元のデータをダウンロードするしかない．場合によっては，さらにその元をたぐりよせ関連するいくつかのデータファイルをダウンロードする必要が生ずる．

2　活用環境

上記のように，e-Stat を本格的に活用するとしても現状では多くの手間と経験が必要となり，ごく限られた専門家が限られた目的で利用するだけに留まらざるをえない．これはなにも官庁データに限らず，さまざまな公開データについても同様で，さらには，各自治体や企業の内部に眠っているデータの活用に際しても言えることである．その場合，一定の枠組みに沿って必要なデータをすべてダウンロードし再構築する方法もあるが，もとのデータの更新などがありうることも考えると，その永続性に疑問が生じる．そこで，e-Stat などから得られるデータを素材とした，ユーザ目線で自由に活用できる一般的な支援環境が必要となる．本報告では，これまで 20 年以上にわたって我々のグループが開発してきた TRAD 環境を例として，必要な条件と実装について議論する．なお，この環境は http:/datascienc.jp/TRAD.html から自由にダウンロードして利用できるよう公開しているので，活用していただければ幸いである．

2.1　InterDatabase

CSV, XLS, RDB などフィジカルなデータ形式の違いを解消し，ネットワーク上に散在したデータに統一的にアクセスできる機能が InterDatabase である．仮想統合データベースと言われるものもこのような機能の実装を目指しているようであるが，次の DandD のようなソフトメディア（媒体）を導入することで汎用性を確保している．

2.2　DandD

DandD は Data and Description の略であり，素材のデータとユーザやソフトウエアを繋ぐソフトメディアの役割を果たす XML ファイル (DandD インスタンス）として実装されている．DandD はフィジカルなデータ形式の違いを吸収するだけでなく，データ表現の違いも吸収する役割を果たす．データ表現の違いにはヘッダー情報から個々の値の違いまでさまざまなレベルの属性の違いが含まれる．DandD は形式的な形式の統一よりも，データの適切な理解を助けることを目標とし，データテーブルを基本にしている．

2.3　TRAD

TRAD は TextilePlot, R and DandD の略であり，適切なデータ解析を効率的に進めるための一般的な支援環境を提供する．これまでの研究成果の集大成として数年前から開発を始めた環境で TextilePlot をデータを視覚的に理解するためのインターフェースとし，データ解析環境 R とのシームレスな連携を実現している．

2.4　DandD ライブラリ

e-Stat で目的のデータがどこに眠っているか調べるのはそう簡単ではない．しかし，DandD インスタンスのライブラリを構築し，ナビゲーション機能を付加することで探索が容易になる．一つの例として 324 本の DandD インスタンスからなる「厚労省患者調査ライブラリ」を公開しているので参考にされたい．