汎用視覚化データ活用環境 TRAD

（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　データ活用環境

社会のさまざまな現場で，蓄積されたものの手つかずのままのデータが急増している．その本格的な活用に踏み込めない原因は，その敷居の高さと人材不足にある．多くの場合，意図的に取得したデータというよりは，業務に伴って副次的に蓄積されたデータであるため，形式はさまざまで扱いは一様ではなく複雑に絡み合っていることが多い．したがって，その全容をきちんと把握しない限り，どのような活用が可能であるか見当すら付けられない．これが敷居の高さである．

データサイエンスが一時ブームだったころ，R が一つのキーソフトウエアとして注目されたが，R は 30 年以上前に開発された S をそのまま継承したデータ解析環境であり，膨大なパッケージを使いこなし与えられたデータを高度に解析する環境としては，今だにその輝きを失ってはいないが，データの流れで言うとかなり下流での使用を念頭に置いたソフトウエア環境である．R のよほどの達人でもないかぎり，データ活用のすべての道のりを R だけでこなすのにはどうしても無理がある．このあたりに， R が一時のブームで終わってしまった一因もあるのではないだろうか．

データの全体像を的確にとらえ理解することで，活用の方向性をなるべく早く見つけだす必要があるデータの上流から中流では，何よりもまず人間の負担が軽いインタフェースで，データ全体を直感的に理解しながら進められる環境が望まれる．その上で，更に詳細な検討や解析が必要な下流になれば，データを R にシームレスに引き継ぎ具体的な活用の方策を固められる環境が望ましい．

本報告では，すでに報告した TexttilePlot や，それをヒューマンインタフェースとする環境 TRAD (TextilePlot, R and DandD) を本格的なデータ活用環境として整備したポイントをいくつか報告する．

2　TRAD

データを数値や記号としてでなく視覚的に捉えるのに，TextilePlot は汎用であり，その中立性と相まってデータ活用環境の基本的なインタフェースとして優れており，データ型を適切に反映することで十分その役目を果たす．しかし，それだけでは効率的なデータブラウジングに十分ではない．TextilePlot 表示の各部の詳細を調整したり， GUI で確かめたりできるだけでなく，気楽に記録や変量の選択・削除・同定・変容などを行え，データ型の変更や名前，コードの変更などができる必要がある．さらに，その変容の過程を記録し，あとで見直したり，再現できる機能も欠かせない．

R との連携はシームレスであるばかりでなく，その役割分担をはっきりさせる必要がある．TRAD では，具体的な数値や記号を確認したり，さまざまなサマリーによって概要をつかんだり，変換を施したりといった作業は R に担わせることで複雑化を避けている．さらに，TRAD では，R の少々時代遅れのヒューマンインターフェイス，貧弱なオブジェクトエディタ，不完全な日本語の扱いなど，データ活用環境としては不満足な点を，独自に作成した GUI で大幅な改善を図った．

さらにデータ活用環境としては，さまざまなデータソースに柔軟に対応できる機能も欠かせない．あらかじめ一つのデータベースにまとめ解析するのも一つの手でありしばしば試みられるが，費用も時間もかかり，ほとんどの場合，活用の方策が見つかる前に息切れしてしまう．TRAD は SQL による RDB へのアクセスばかりでなく， CSV や Excel ファイルなどをドラッグドロップするだけで読み込むことができるように作られている．しかし，データファイは RDB のように形式が整っているとは限らず，さまざまな状況に対応できる柔軟できるヒューマンインターフェースが必要となる．TRAD ではこれを Sytanx と Semantics の 2 段階で実現している．第 1 段階では，データファイルをどう読み込むかだけを指定し，第 2 段階で読み込んだデータの Syntax を補正することで，柔軟な対応を可能にしている．

参照: http://datascience.jp/TRAD.html