基般

情報教育と統計教育 No. 2

手順的な自動処理と機械可読データ

奥村晴彦
三重大学教育学部

本会情報処理教育委員会の「日本の情報教育・情報処理教育に関する提言 2005」[1] では，情報教育における「手順的な自動処理」の体験が重要であるとしている．実際，「ちょっと頭を使えば複雑な作業も自動化できる」というセンスをみんなが共有すれば，世の中はずっと楽になるのではないか．

ところが，我々の身の回りには，せっかく Excel で作られながら，自動処理を念頭においていないデータが蔓延している．Excel を教えている教育現場に問題はないのだろうか．

ネ申（かみがみ）ネ申の黄昏（たそがれ）

Excel ファイルの帳票類を埋める作業でうんざりしておられる読者は多いと思う（**図 -1**）．

たとえば「該当のものを○で囲む」という場合，挿入→図形で○を描くべきか．Mac・Windows 間でレイアウトが崩れて○の位置がずれないか．該当選択肢を枠線で囲むか，頭に文字「○」を強引に挿入するのはどうか．□にチェックを付けよという指示であれば，図形で✓を描くのか，□を■または☑という文字に置換するのか．

フリガナ		性別	生年月日	年 月 日（ 才）	部局名		
氏名		男・女			職名		
職務内容（該当箇所に○をつける）	1. 医療事務	2. 医療技術	3. 図書業務	4. 情報処理・技術			
	5. 一般事務	6. 教員	7. その他（ ）				
VDT作業の有無	有 ・ 無	「有」の方のみ以下の質問にお答えください。					
病歴	眼圧が高いと診断されたことがありますか		VDT作業歴	年 月 ～ 年 月			
	有	（ 年頃） 無					
	緑内障と診断されたことがありますか						
	有	（ 年頃） 無					
VDT作業の状況	(注2) VDT作業区分（該当区分に○をつける）	A		(注2) 作業の種類（該当の種類に○をつける）	1. 単純入力型		
		B			2. 拘束型		
		C			3. 対話型		
	1回の平均VDT従事時間数	時間 分			4. 技術型		
	1日の平均VDT従事時間数	時間 分			5. 監視型		
	1週間の平均VDT従事日数	日			6. その他の型		
項目		有無・適否		項目		有無・適否	
目が疲れる		有	無	頭が重い（痛い）		有	無
目が痛い		有	無	肩や腕がこる		有	無
目がかすむ		有	無	手指の力が入りにくい		有	無

図 -1　おそらく手集計される Excel ファイル

いずれにしても，これでは自動集計は無理なので，受け取った人は１つ１つ Excel ファイルを開いて，画面を見ながら，あるいはいったん紙に印刷して，手動で集計しているのであろう．

しばしばセルはすべて正方形に揃えて「Excel 方眼紙」にする．セルを結合し，罫線を引いて，複雑な帳票を作る．Facebook グループ「エクセル方眼紙根絶委員会」には，この種の実例がたくさん報告されている．

このように，せっかく電子化しても，紙の帳票文化をそのまま持ち込むため，「神（紙）Excel」あるいはネットスラングでいう「ネ申 Excel」になってしまう[2]．構造化されておらず，入力は面倒だし，アクセシブルでないし，データとしての再利用は困難である．

Excel の「オープンデータ」

オープンデータを提供するはずの政府統計の総合窓口 e-Stat[☆1] にある Excel ファイルも「ネ申 Excel」が多い．

e-Stat のデータの URL は長く，永続性もなさそ

☆1　http://www.e-stat.go.jp/

図 -2　e-Stat の労働力調査データ

うなので，URL は示さないが，**図 -2** のように労働力調査→基本集計全都道府県→長期時系列データとたどったページの最初のファイルを開いてみよう．

最初のシートの最初の部分を**図 -3** に示す．

プロプライエタリな XLS 形式を使っているのも問題であるが，テキスト形式に直すツールが存在するし，前回解説した言語「R」にも XLS 形式・XLSX 形式のファイルを読むライブラリが存在する（例：readxl）．

真の問題は，論理的な構造が明瞭でなく，セル結合と罫線による視覚的な構造に頼っているので，何年何月の労働力人口がどこのセルに入っているかは，人間が目視しないと分かりづらいことである．CSV 形式で保存してテキストエディタで見てみれば，問題点がよく分かるであろう．

機械可読な表形式のデータ

上で例として挙げた労働力人口データは，**表 -1** のような形式であれば，人間にもコンピュータにも判読しやすい．「男女計」は，「男」と「女」から導けるなら，不要である．

このような表形式データを表すには，Excel ファイルでなくても，CSV やタブ区切りのテキストで十分である．なお，この表が何を意味しているかは，別にメタデータとして与えておく必要がある．

このように，セル結合をせず，1 行目に列名（変数名）を与え，2 行目以降に値を並べるのが，標準的な表形式データの表し方である．年月は年と月

	A	B	C	D	E	F	G
1	[基本集計]				長期時系列表１　ａ－１　主要項目（労働力人口・就業者・雇		
2	[Basic Tabulation]				Historical data 1 a-1 Major items (Labour force, Employed person, Employee, Ur		
3							
4					（万人）	(ten thousand persons)	
5	年　月 Year and month				季節調整値	Seasonally adjusted series	
6					労働力人口		
7					Labour force		
8					男女計	男	女
9					Both sexes	Male	Female
11	昭和28年	1月	Jan.	*	(4122)	(2416)	(1708
12	1953	2月	Feb.	*	(4001)	(2375)	(1625
13		3月	Mar.	*	(4008)	(2382)	(1624
14	※注_Notes	4月	Apr.	*	3956	2369	1588
15		5月	May	*	3917	2357	1562
16		6月	June	*	3940	2358	1584
17		7月	July	*	3957	2367	1592
18		8月	Aug.	*	3963	2361	1602
19		9月	Sept.	*	3982	2369	1613
20		10月	Oct.	*	4015	2383	1633
21		11月	Nov.	*	4018	2381	1637
22		12月	Dec.	*	4017	2383	1633
23	昭和29年	1月	Jan.	*	3959	2364	1592
24	1954	2月	Feb.	*	3996	2379	1615
25		3月	Mar.	*	4021	2387	1633
26		4月	Apr.	*	4036	2388	1648
27		5月	May	*	4073	2404	1668
28		6月	June	*	4039	2396	1643
29		7月	July	*	4087	2413	1674
30		8月	Aug.	*	4088	2422	1666
31		9月	Sept.	*	4079	2425	1655

図 -3　e-Stat の労働力調査データ

年月	男	女
1953-01	2416	1708
1953-02	2375	1625
1953-03	2382	1624
1953-04	2369	1588
……	……	……

◀ 表 -1　日本の労働力人口の季節調整値（万人）

氏名	性別	好きな数
奥村晴彦	男	3.1416

表 -2　Excel ファイルによるアンケートの例

に分けてもよいが，YYYY-MM や YYYY-MM-DD や YYYY-MM-DD hh:mm:ss のような Excel にも理解できる形式なら，1 つのセルに入れるほうが楽である（R でも大丈夫である）．

仮にアンケートを Excel で集めて集計する場合にも，**表 -2** のような標準的な形式であれば，簡単に自動集計できる．このような XLS/XLSX ファイルがいくつかカレントディレクトリに置いてあるとして，R ならば，たとえば

```
library(readxl)
f = dir(pattern="*.xls*")
x = do.call(rbind,
        lapply(f, read_excel))
```

で全データを縦に結合したデータフレーム（行列のようなもの）ができる．これを Excel で開きたければ，

```
write.csv(x, "x.csv",
      fileEncoding="SJIS")
```

として 1 つの CSV ファイルに収めればよい．文字

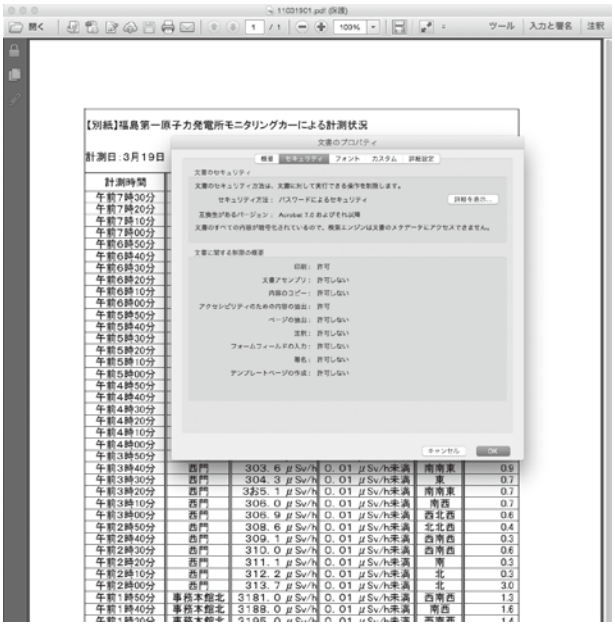

図 -4　東電の 2011 年 3 月 19 日の敷地内線量測定値（発表時の URL☆2）．数値を全角文字で単位付きで入力するのは余計な手間がかかるだけでなく，午前 3 時 20 分の「3お 5．1 μ Sv/h」のように入力を間違えても気づきにくい．なお，ファイルは保護設定されており，通常の方法ではコピー・テキスト抽出できない．

コードは UTF-8 にしたいところだが，日本語環境の Excel はシフト JIS の CSV しか開けない．

Excel で CSV ファイルを読み込む際には，"1/4" のような分数や "Oct4" のような遺伝子名を日付と誤解する問題に注意しなければならない．

東日本大震災と機械可読データ

東日本大震災では，大量の重要な「データ」が公開されたが，その多くは，データとして再利用することが難しい PDF ファイルであった．

阪神淡路大震災のときと違って，スマホなどの情報機器，Twitter などの通信手段が活躍した[3)～5)]．一方で，国や自治体，東京電力（以下，東電）の情報交換ツールはいまだにファクスであった．東電の計画停電情報が当初はファクスのスキャン画像で発表されたが，アクセス過多でつながりにくくなり，ボランティアがミラーしたりテキストに起こしたりした．

東電は地震発生時刻まで福島第一原子力発電所敷地内での放射線量を Web でリアルタイム公開していたが，事故後は手動でデータ公開した．数値はほぼ全角文字で３０５．１μ Sv/h のように単位も付けて Excel に入力され，PDF で公開された．最初のうちは保護設定もされていた．ときどき「.」と「,」を間違えたり，「μ」を抜かしたり，「0」が「お」になったり，時刻の午前と午後を間違えたりしていた（**図 -4**）．私はこれをデータ化するのにたいへん苦労した．

事故でたいへんな時期に，きれいに入力して公開を続けられたことは評価するが，単位を付けずに数値だけ半角で入力するほうがずっと楽であったであろうし，グラフを描くなどデータとして活用できる．時刻も「午前 3 時 20 分」ではなく「2011-03-19 03:20」のような日時にすれば，簡単な操作で一定間隔の値が入力でき，別の場所から日付を拾ってくる手間が省け，時系列グラフが描きやすい．

文部科学省（以下，文科省）は 47 都道府県の毎時の放射線量を PDF で毎日 2 回公開した．これもデータとして使えるものにするために苦労した．現在は原子力規制委員会が CSV・JSON 形式でリアルタイム公開している．

厚生労働省（以下，厚労省）は，自治体から送られてくる大量の食品中の放射性物質の検査結果の Excel データを手作業で貼り合わせ，休日を除く毎日，PDF で公開した（**図 -5**，現在は PDF で週 1 回，PDF と Excel ファイルで月 1 回公開）．自治体も厚労省も手作業のため，しばしば重複や記入ミスが生じていた．PDF では検索も困難なので，（公財）食品流通構造改善促進機構がボランティアでこれをデータベース化し，検索可能な形で再公開していた[6)]．そのサイトが 2012 年 4 月に終了し，全データとソースコードがで公開された☆3．私はこの後を引き継ぐ形で，PDF を毎日手作業（スクリプトで変換，エラーが生じたら手動で修正）でデータ化し，CSV 形式で再公開し，簡単な検索インタフェースも付けた☆4．2013 年 11 月からは，国立

☆2　http://www.tepco.co.jp/nu/monitoring/11031901.pdf
☆3　https://github.com/udawtr/yasaikensa
☆4　http://oku.edu.mie-u.ac.jp/food/

保健医療科学院が厚労省の委託でデータベース化を引き継いだ.

厚労省の PDF ファイルは毎夜遅く公開された. 遅くまで仕事をされていることには感謝するが，最初から発生源により近いところでデータベースに入力できる仕組みを構築するか，あるいは少なくとも自治体からの Excel データを自動でデータベースに登録できる仕組みを構築しておけば，手作業による間違いもなく，皆が楽をできたのではないか.

どうすればよいか

冒頭に挙げた「日本の情報教育・情報処理教育に関する提言 2005」[1] で論じられているような「手順的な自動処理」のセンスを学んでいれば，たとえ自分でプログラムを書けなくても，自動処理を考慮したデータの集め方・公開の仕方があることを，想像できたかもしれない.

さらに積極的に，身近なデータを実際に処理してみる演習を，もっと情報教育に取り入れてはどうであろうか.

参考文献

1) 日本の情報教育・情報処理教育に関する提言 2005（2006.11 改訂／追補版），情報処理学会情報処理教育委員会，http://www.ipsj.or.jp/12kyoiku/teigen/v81teigen-rev1a.html
2) 奥村晴彦：「ネ申 Excel」問題，情報処理学会情報教育シンポジウム SSS2011 論文集，pp.93–98 (2013).
3) 奥村晴彦：震災とソーシャルネットワーク，情報処理，Vol.52, No.9, pp.1072–1073 (Sep. 2011).
4) 奥村晴彦，辰己丈夫，藤間　真：大震災で見えてきた情報教育の課題，情報処理学会情報教育シンポジウム SSS2011 論文集，pp.25–32 (Aug. 2011).
5) Okumura, H. : The 3.11 Disaster and Data, Journal of Information Processing, Vol.22, No.4 (2014).
6) 杉山純一，宇田　渉：安全・安心のための食の情報処理，情報処理，Vol.52, No.11, pp.1370–1375 (Nov. 2011).

（2015 年 4 月 30 日受付）

検査法（Ge／NaI）	採取日（購入日）	結果判明日	結果（Bq/kg） ヨウ素−131	セシウム−134	セシウム−137
Ge	H24.3.22	H24.3.22	< 5	< 5	< 5
Ge	H24.3.22	H24.3.22	< 5	< 5	< 5
Ge	H24.3.21	H24.3.21	< 5	< 5	< 5
Ge	H24.3.21	H24.3.21	< 5	< 5	< 5
Ge	H24.3.14	H24.3.22	< 12	105	161
Ge	H24.3.14	H24.3.22	< 10	67	137
Ge	H24.3.14	H24.3.22	< 9	31	172
Ge	H24.3.21	H24.3.22	< 4.7	< 4.9	< 4.9
Ge	H24.3.21	H24.3.22	< 4.1	< 4.3	< 5.0
Ge	H24.3.21	H24.3.22	< 3.9	< 4.6	< 4.5
Ge	H24.3.21	H24.3.22	< 1.9	< 1.8	< 1.4
Ge	H24.3.22	H24.3.22	< 5	< 5	
Ge	H24.3.21	H24.3.21	< 2	< 3	< 2
NaI	H24.3.19	H24.3.20	−	< 25	
NaI	H24.3.19	H24.3.20	−	< 25	
NaI	H24.3.19	H24.3.20	−	< 25	

図 -5　厚労省「食品中の放射性物質の検査結果について」第 350 報（2012 年 3 月 22 日）の一部．個々のセシウムの値が不明な場合にセル結合を使うという難点は，その後「セシウム合計」欄を追加することで解決された．一方で，検出限界 x で不検出であることを意味する "< x" という記法は，1 つのセルに 2 つの情報を入れることになり，しばしば "<" を見落として誤解を生じる原因にもなったが，改善に至っていない.

奥村晴彦（正会員） okumura@okumuralab.org

三重大学教育学部教授（情報教育）．LHA データ圧縮アルゴリズムの開発者，『LaTeX2e 美文書作成入門』（現在第 6 版）の著者.