Data processing and analysis with R at OECD

Scope for A/S Open Source Software in an International Organisation

Bo Werth
Statistician OECD STI/EAS

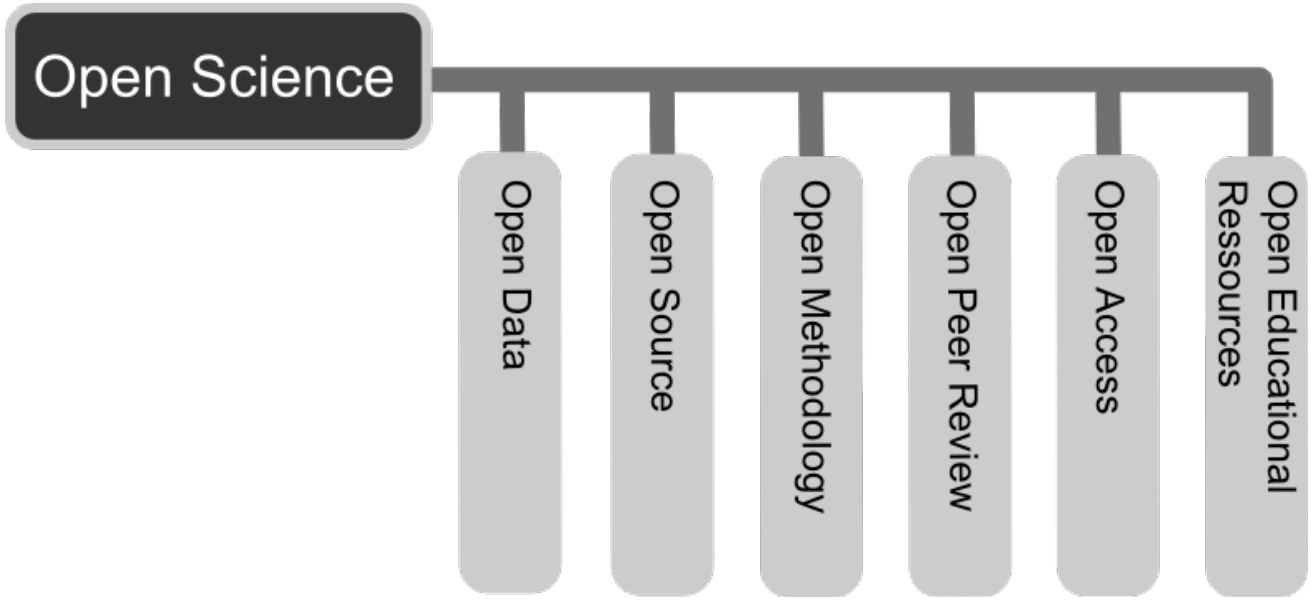
Disclaimer

The information contained in this presentation has not been reviewed by the Organisation and do not necessarily represent the official views of the Organisation or of the governments of its member countries.

About OECD

- "Conventional Wisdom Central" krugman.blogs.nytimes.com
- evidence-based reports for policy recommendation
- exchange platform for government officials
- directorates covering different aspects of the economy
  - education, employment, international trade, tax regulations, environment, financial sector, science and technology...
- expert group meetings take place twice per year
- considerable overall budget
- allocated to deliverables every 2 years
- relatively small IT budget (6-7%)

The six principals of open science



Source: Wikipedia

Rogoff and Reinhart: Growth in a Time of Debt

When a country's debt-to-GDP level gets above 90%, real GDP growth takes a big hit

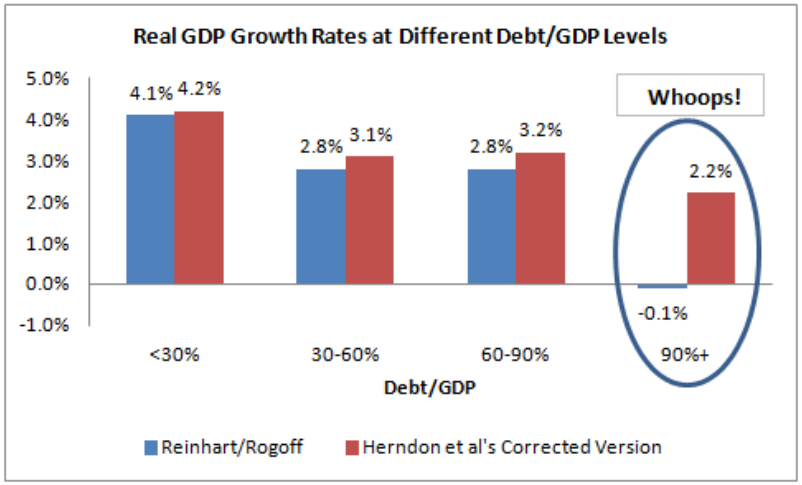
Businessweek: FAQ: Reinhart, Rogoff, and the Excel Error That Changed History

The Atlantic: Forget Excel: This Was Reinhart and Rogoff's Biggest Mistake

Carmen M. Reinhart & Kenneth S. Rogoff, 2010. "Growth in a Time of Debt," American Economic Review, American Economic Association, vol. 100(2), pages 573-78, May.

Herndon, Ash, and Pollin

They replicate R&R's original work and make various corrections to a) methods and data choices and b) a "spreadsheet error," the latter where R&R appear to have left out some important data that has a big impact on their results.



Source: http://jaredbernsteinblog.com/not-to-pile-on-but-correcting-reinhart-and-rogoff/

Peer-review: Mail Avengers

A paper that largely consists of the words "Get me off your *** mailing list" repeated 863 times was accepted by a journal that claims to be peer reviewed.

Get me off Your Mailing List

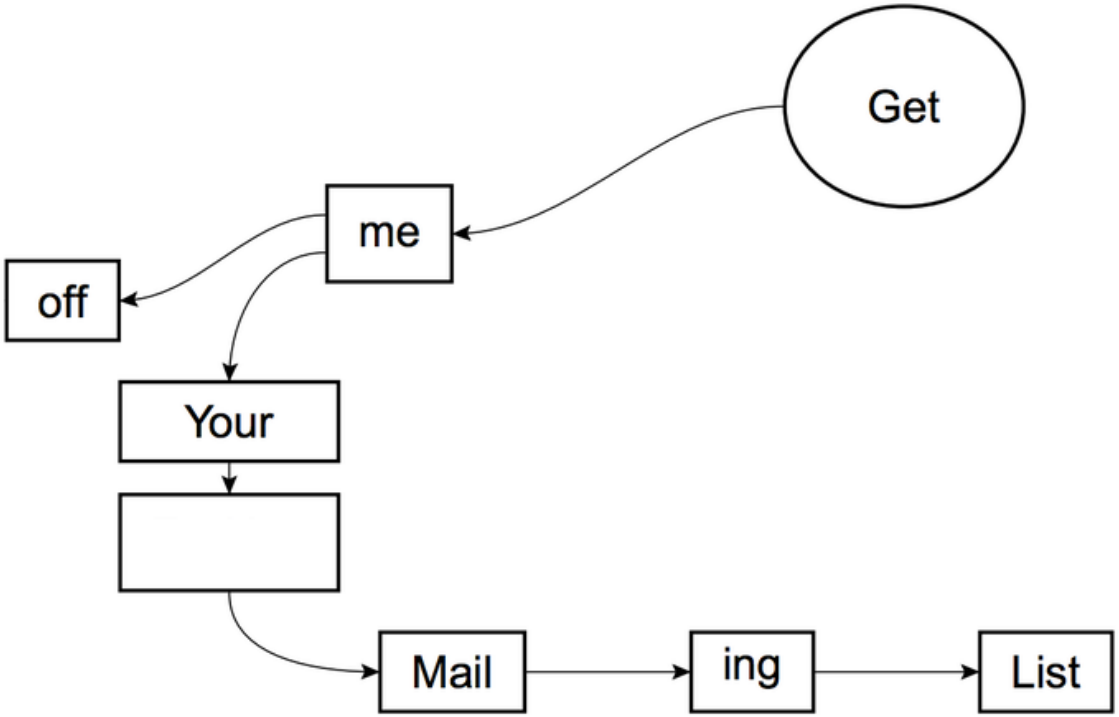
David Mazières and Eddie Kohler
New York University
University of California, Los Angeles
http://www.mailavenger.org/

Abstract

Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list. Get me off your mailing list.

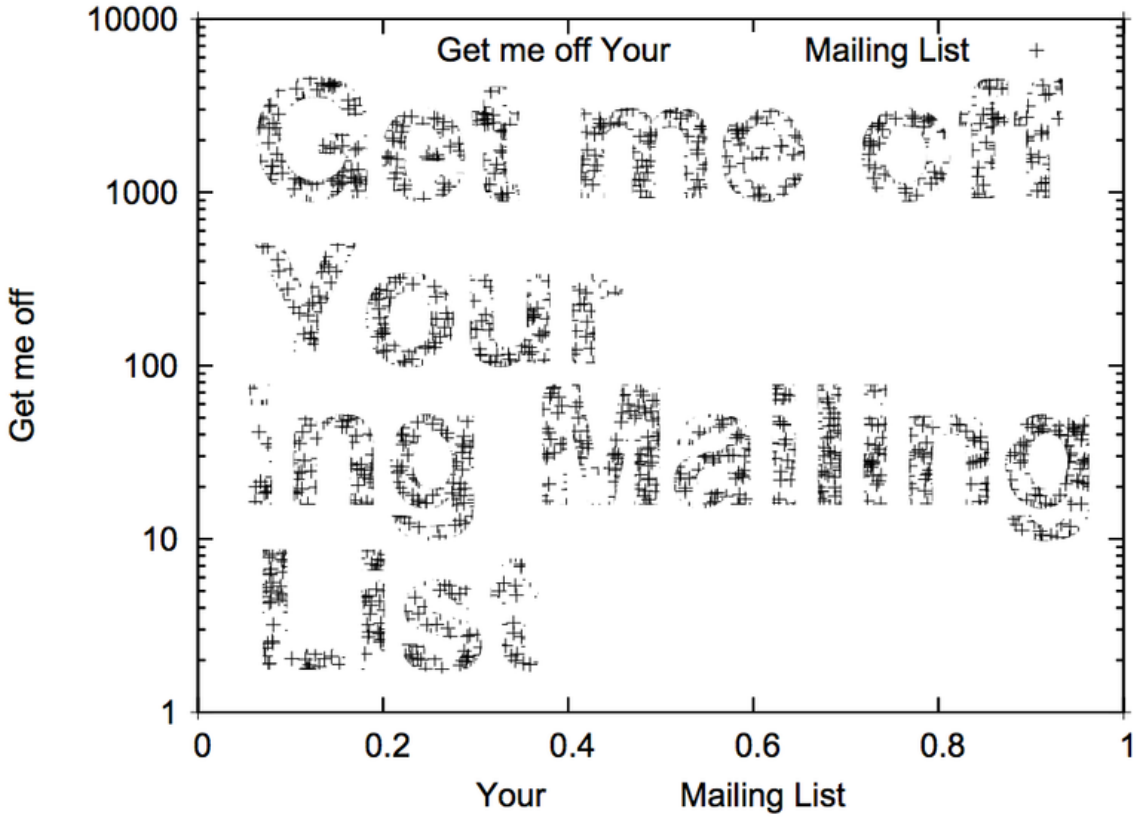
Source: http://www.scs.stanford.edu/~dm/home/papers/remove.pdf

Mail Avengers: Figure 1



Source: http://www.scs.stanford.edu/~dm/home/papers/remove.pdf

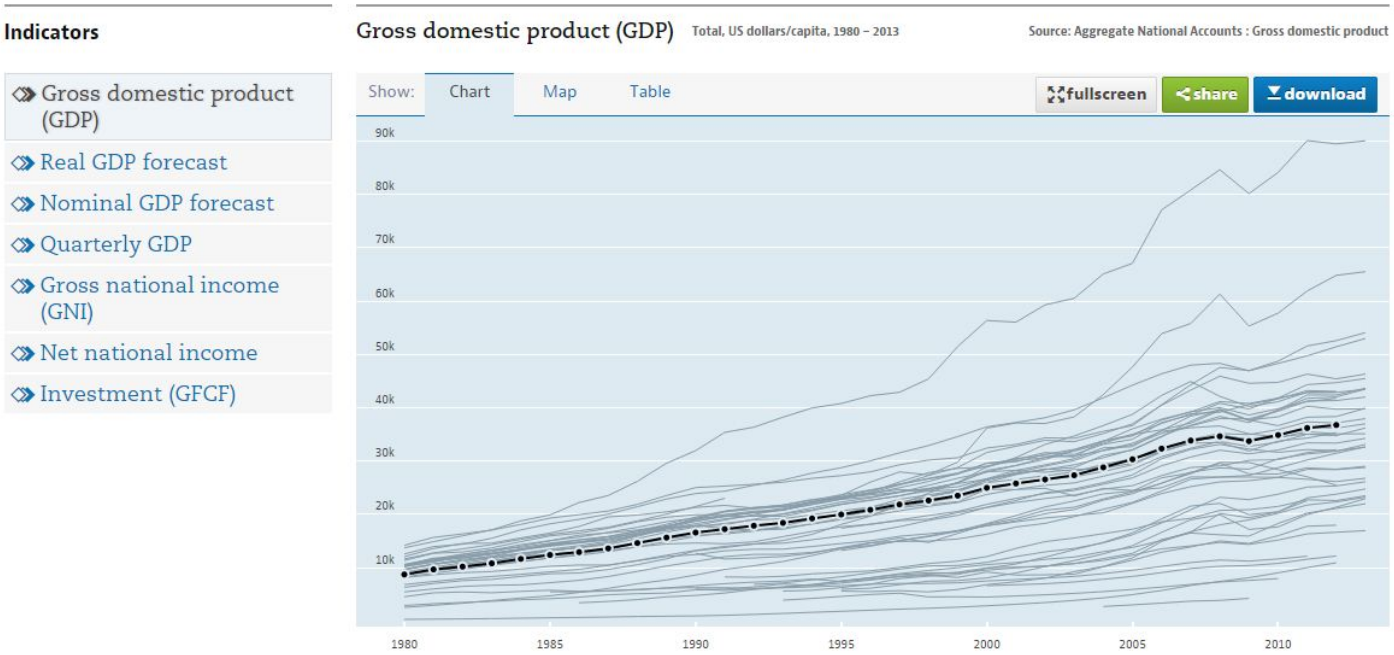
Mail Avengers: Figure 2



Source: http://www.scs.stanford.edu/~dm/home/papers/remove.pdf

OECD Data Portal

Increase free accessibility to OECD's publications and data, while increasing dissemination and maintaining a sustainable and effective publishing operation



http://data.oecd.org/

OECD Open Data

OpenDataAPI http://stats.oecd.org/opendataapi/

Statistical Data and Metadata eXchange (SDMX)

Information about practical implementations: http://sdmx.org/

Publication Workflow

Statisticians

- identifying and harmonising data
- graphics for analysis and publication

Analysts

- narrative for draft report

Senior Analysts

- editing for final report

Publishing and communications

- publishing on OECD iLibrary

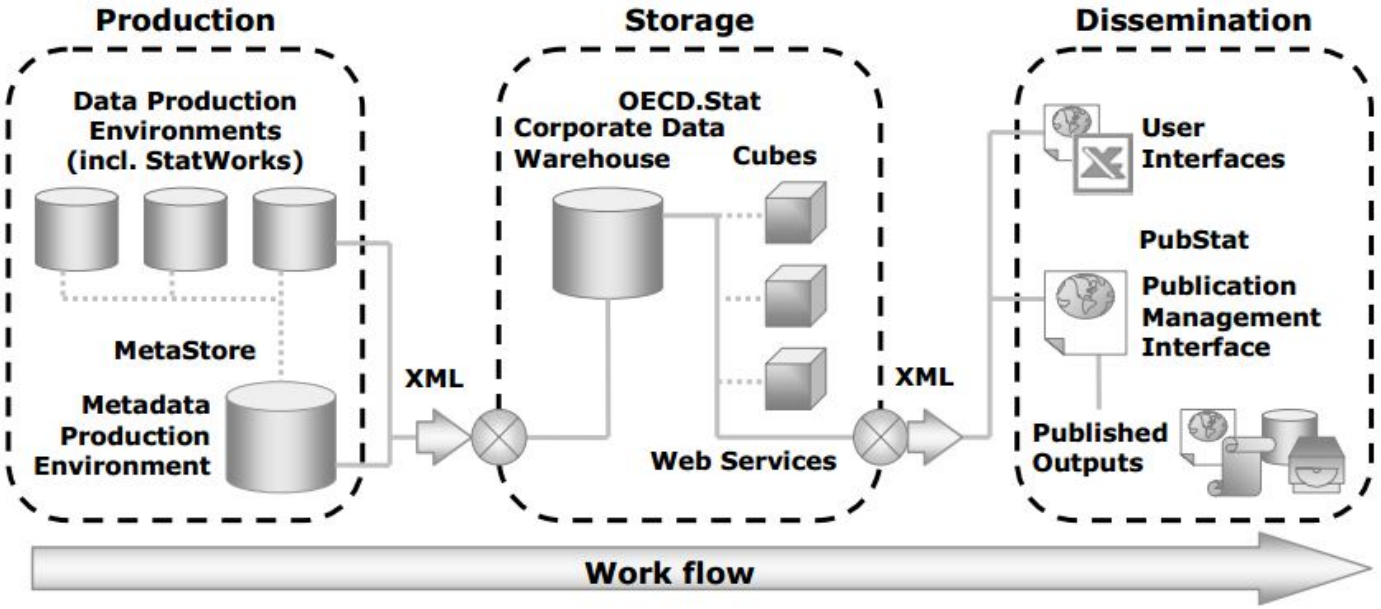
Licensed A/S Software

- EViews: time-series oriented econometric analysis
- FAME: Forecasting Analysis and Modeling Environment
- GAMS: high-level modeling system for mathematical programming and optimization
- MATLAB: numerical computing environment for matrix manipulations
- Prognoz: BI, analytics and visual discovery solutions
- SAS: Business Analytics and Business Intelligence Software
- Stata: Data Analysis and Statistical Software
- Tableau: Business Intelligence and Analytics
- TROLL: integrated software system for econometric modelling and statistical analysis

=> Large variety and high license cost

In-house developments

- StatWorks: SQL-based information system for the management of statistical production data



- Implemented in 2004, maintained and enhanced by the IT department
- oriented at statistics, limited analytical functionality

Infrastructure with Open Source Software

Requirements

- managed by staff with varying IT expertise
- interface with other languages and software
- front-end capacity
- flexible and intuitive chart libraries
- outputs for print and online publishing

Candidates

Front-end: web browser integration

Motivation

- make usage as simple and intuitive as possible
- increase accessibility to developed procedures and back-ends
- ultimately consider public server hosting

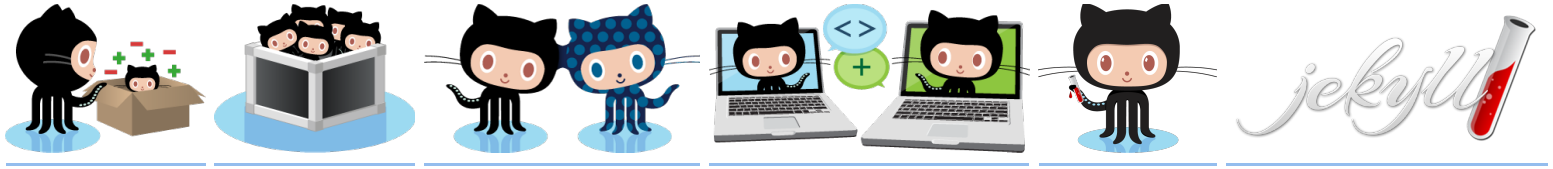
Candidates

Back-end: source code sharing

Motivation

- benefit from OSS user community developments
- systematic user-driven enhancements
- co-operation and version control
- testing, transparency and issue tracking
- wiki documentation, jekyll project sites

Github



Reporting

Requirements

- efficient templating syntax
- markdown support
- flexible outputs with Pandoc universal document converter

R Implementations

rapport: R templating system based on pander

ReporteRs: generate Microsoft Word, Microsoft PowerPoint and HTML reports

Visualisation

R

- ggplot2: R plotting system
- networkD3: D3 network graphs with R htmlwidgets
- rCharts: R interface to JavaScript libraries

Python

- Bokeh: Python interactive visualization library
- d3py: Python interface to build interactive, javascript based plots
- matplotlib: Python 2D plotting library

Julia

- Gadfly: plotting and visualization system for Julia

Online Publishing

TERMINALFOUR: current platform for website authoring

- websites must be tracked with their IDs
- media content items must be uploaded individually
- management of resources (js, css) requires admin intervention

Alternative: Static Blog Generation

- file-based + database-free site generators (overview: modernstatic, staticgen)
- e.g. jekyll + octopress (ruby), pelican (python)


Standards for OSS Usage

- appropriate licensing
- categorize into users and contributors
- formalize style, documentation and testing requirements
- specialised training offers (plot creation, function documentation, packaging, testing etc.)
- exchange about best practicies with national statistics offices
  - Italian National Institute of Statistics (ISTAT): Use of R in Business Surveys (pdf)
  - Statistics Netherlands (CBS): The Introduction and Use of R Software (pdf)
  - UK Office for National Statistics (ONS) Use of R in the UK ONS (pdf)

SDMX Java library

Java functions developed by Attilio Mattioc at Bank of Italy

- supported clients: ECB, Eurostat, IMF, ILO, OECD, INEGI (see list on github)
- R package with connector functions: RJSDMX

Cloud Hosting

ICIO Foreign Demand Domestic Value Added

indicator calculation platform (shiny server + Azure): http://oecd-icio.cloudapp.net:3838/

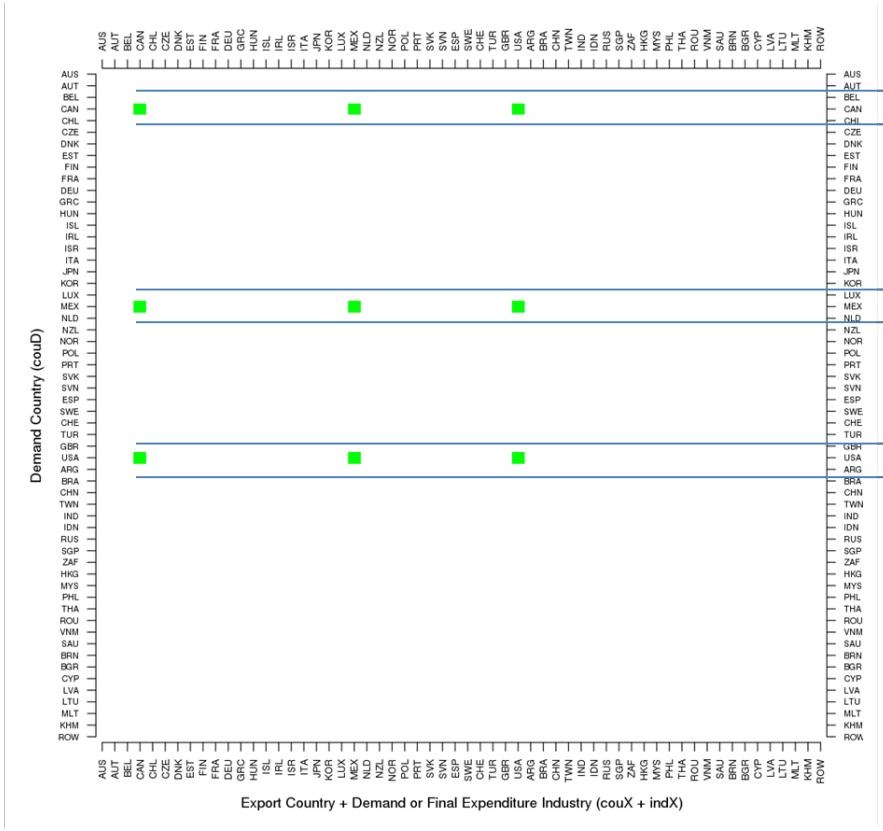
- select dimensions
- subset data from multidimensional arrays
- perform calculations:

```
data.couX.indX <- data.conv1 * data.demand
aaa <- xB %*% data.couX.indX
aaa <- apply(aaa, 1, sum)
```

- aggregate and display results (table, barchart, map)

ICIO Foreign Demand Domestic: couD, couX, indX

“NAFTA”



Canada

Mexico

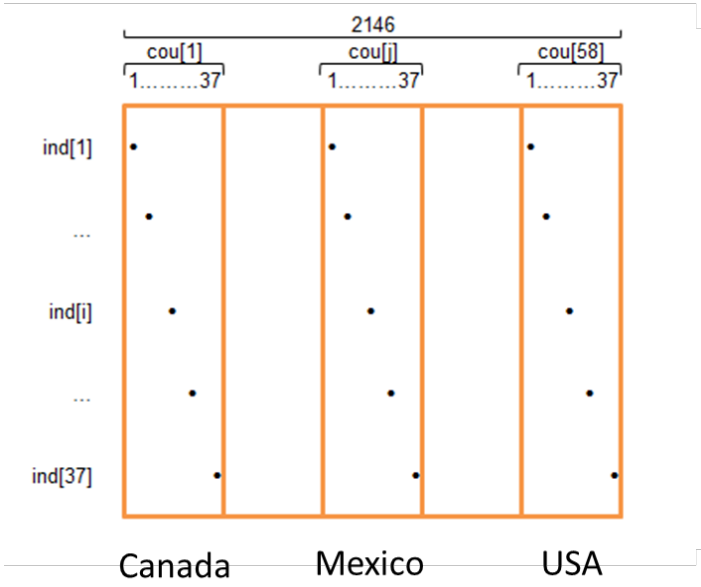
USA

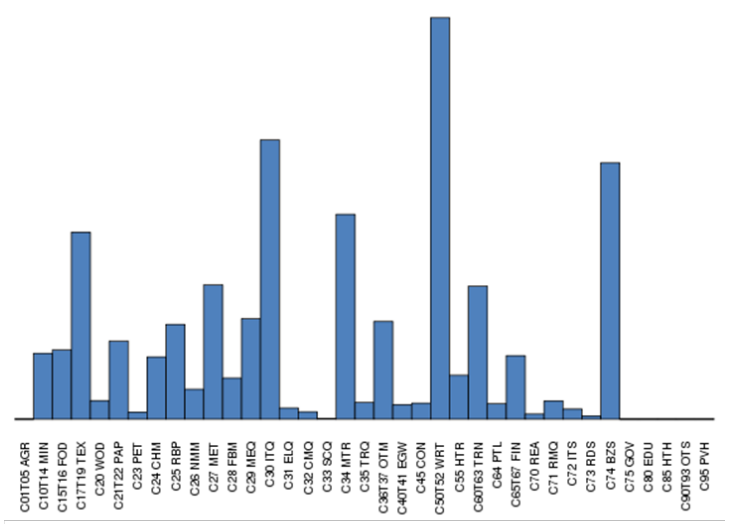
ICIO Foreign Demand Domestic Value Added: calculation + aggregation

res.matrix = xB %*% data

res.vector = apply(res, 1, sum)

```
$ sparse demand Matrix
            2146      58
$ xB: coefficient Matrix
            2146    2146
```





ICIO Foreign Demand Domestic Value Added: aceEditor

Console input

```
1  list(conv=dim(conv),
2  demand=dim(demand),
3  xB=dim(xB),
4  aaa=length(aaa)
5  ##, data.by.ind=length(data.by.ind)
6  ##, data.by.cou=length(data.by.cou)
7  )
```

Calculation formula

```
1  data.couX.indX <- conv * demand
2  aaa <- xB %*% data.couX.indX
3  aaa <- apply(aaa, 1, sum)
```

Evaluate

Coefficients:

eB: Employment

Year:

2009

Demand concept: (demand)

GRTR: Gross Trade

Demand or Final Expenditure Industry: (indX)

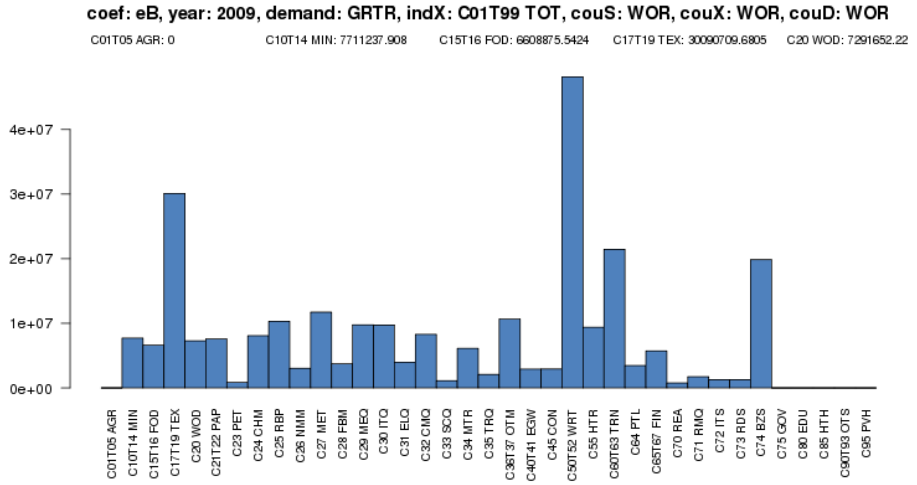
Plot: Bar: Ind | Table: Data: Ind | Plot: Bar: Cou | Table: Data: Cou

```
$conv
[1] 2146   58

$demand
[1] 2146   58

$xB
[1] 2146 2146

$aaa
[1] 2146
```



References

- ProgrammableWeb link to OECD Open Data API
- European Union Open Data Portal
- rOpenSci: Transforming science through open data

Contact

https://github.com/bowerth
bo.werth@gmail.com
bo.werth@oecd.org
OECD Directorate for Science, Technology and Innovation
2, rue André Pascal
75775 Paris Cedex 16