European Public Sector Information Platform

Topic Report No. 2012 / 06

Data Wrangling

Authors: Daniel Dietrich and Henri Myrttinen

Published: May 2012

Keywords

Open Data, Data, Wrangling, Tools, PSI, re-use,

1 Executive Summary

Data wrangling, the processes of searching, finding, retrieving, cleaning, formatting, analysing, and presenting data, is becoming increasingly central to individuals and institutions dealing with ever-increasing information flows. Given both its quality and quantity, the opening of PSI data for re-use both necessitates the use of data wrangling tools and allows for new, innovative solutions to be created through these processes, e.g. through combining and mashing up previously isolated, compartmentalised datasets. These can ideally help increase citizen participation, improve PSB service delivery, increase transparency and good governance as well as open up new commercial possibilities. Due to the increasing centrality of complex data sets and information flows to all aspects of life, from the mundane checking of bus timetables to coordinating the fight against global pandemics, managing PSI through improved data wrangling is not merely a nice luxury but rather an inescapable necessity.

2 Introduction

Data Wrangling is a colloquial term for finding innovative ways for the increasing challenges of working with vast amounts of data. With the technical advances in ICT over the past decades, both the quality and quantity of data which can and are being produced and accessed daily are incomparably larger than previously - and the trend is accelerating. The data comes from a variety of sources, such as private individuals, corporations, organisations and also Public Sector Bodies (PSBs). In the case of the latter, the opening up of public sector information (PSI) has the potential of making some of the most interesting and valuable sources of data for available for re-users. At the same time, however, the sheer quantity and variety of formats of PSI data makes its re-use challenging.

The exponential growth of the quantity of data and its increased availability pose increasingly pressing questions of how to deal with this material. With the rise of open data and big data, working with data becomes increasingly central for different actors in society, such as governments, PSBs, research and educational institutions, media, civil society organisations, advocacy groups, individual citizens and last but not least businesses. Working with data effectively requires new skills and tools, as this data needs to be accessed, sorted, analysed and put into a format in which it is useable and understandable. This is especially the case if the data is “messy” and comes in heterogeneous formats, which can often be the case with PSI. Here, new data wrangling technologies offer an increasing range of possibilities for managing and making sense of the data sets.

According to a 2011 McKinsey study[1], however, there will be skills shortage in data expertise about 50-60% by 2018 in the US alone. Consequently there is a need for training

[1] McKinsey Global Institute, 2011. “Big data: The next frontier for innovation, competition, and productivity”

facilities and courses to work with emerging tools. In part, the opening up PSI data for re-users outside of PSBs can also help fill this gap, as the re-users may be able to find more effective and efficient ways of 'wrangling' data than the original users might have been able to.

Data wrangling or management covers a range of ways of working with data, such as data management, data analysis, data mining, and data visualisation. Ultimately, it aims at gathering, organising and managing increasingly large and complex sets of data, thus increasing the efficiency with which this date can be used and re-used.

3 Market: Who needs Data wranglers?

As dealing with growing amounts of data efficiently and speedily becomes increasingly central to contemporary life, the need for data wrangling is virtually unlimited. The management of complex data sets is no longer merely the purview of specialists directly working with them, e.g. in administration or research. Rather, they are now part of everyday activities such as checking the metro connections on one's smartphone to see if one can still make it to the cinema in time.

While many of these datasets are, on their own, often only of interest to a narrow group of users focusing on the particular issue, the opening up of data allows for "mashing up" and combining previously isolated datasets into new products and services which can turn out to be highly relevant to everyday lives of citizens or of high commercial value.

In terms of PSI re-use, key potential beneficiaries are any individuals or organisations which need to manage, compile, analyse or in other ways use increasingly large and complex sets of data. These potential beneficiaries include:

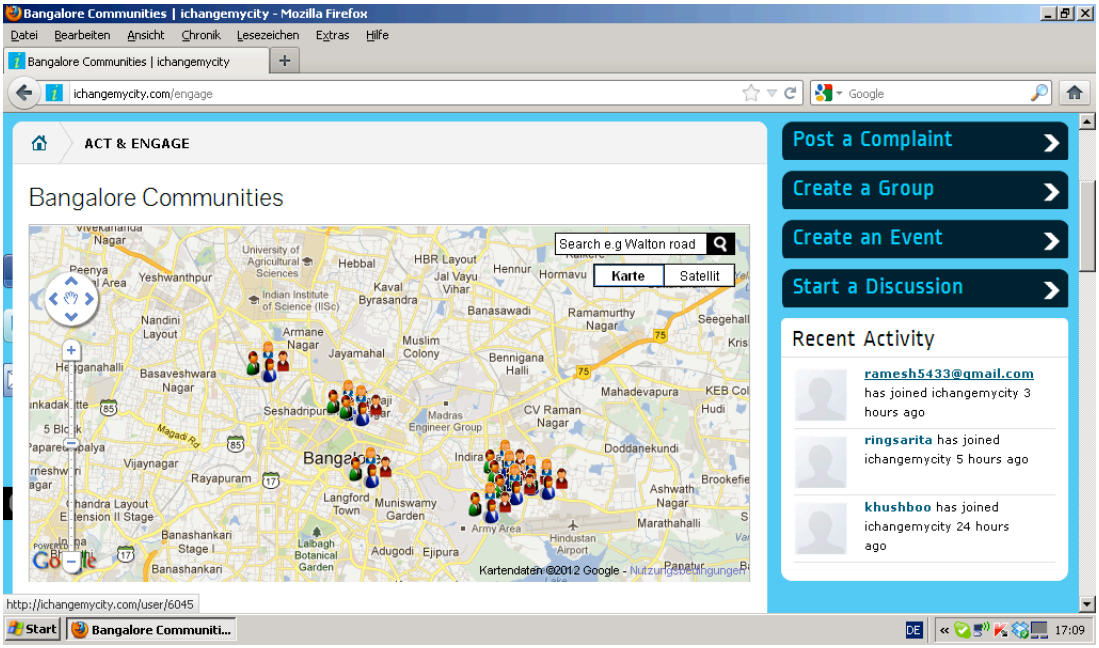
- the PSBs themselves
- Citizens
- Media
- research and educational institutions
- NGOs, Think Tanks and other Orgs
- and commercial enterprises

PSI is unique both in terms of its breadth and depth, and also the fact that it is mostly produced by highly specialised institutions and individuals. PSI spans a vast range of fields from geodata to parliamentarians' attendance records; from development aid budgets to the planned reparation of potholes; from records of ancient artefacts to bus timetables. The data sets have historically often been isolated from each other due to bureaucratic compartmentalisation. By opening PSI up to re-use, the various beneficiaries can potentially benefit from synergies and creative approaches to data wrangling which would not have come about without the opening up of the data and the subsequent re-combinations that this allows.

4 Examples of Successful Data Wrangling

1. Increased citizen engagement

The Bangalore-based website "I Change My City"[2] combines PSI such as land use regulations, geodata, locations and contacts of PSBs and elected representatives with possibilities of citizens to post complaints, take part in discussion forums, contact other citizens and form civic action groups. The site has been set up and is run by the Bangalore-based non-profit Janaagraha Centre for Citizenship and Democracy



The site allows citizens to engage with PSBs and elected officials in pushing for better and more efficient service delivery. The service can be accessed by smartphone via an app.

Linked to the I Change My City site is also the anti-corruption site "I Paid A Bribe."

2. Improved CSO advocacy work

Tactical Technology[3] offers a variety of data wrangling tools to CSOs for data-based advocacy. The organisation has worked with a large range of CSOs on a variety of issues ranging from land-grabbing to maternal health and slavery to environmentally damaging investments. Much of the data used for the advocacy work is open PSI.

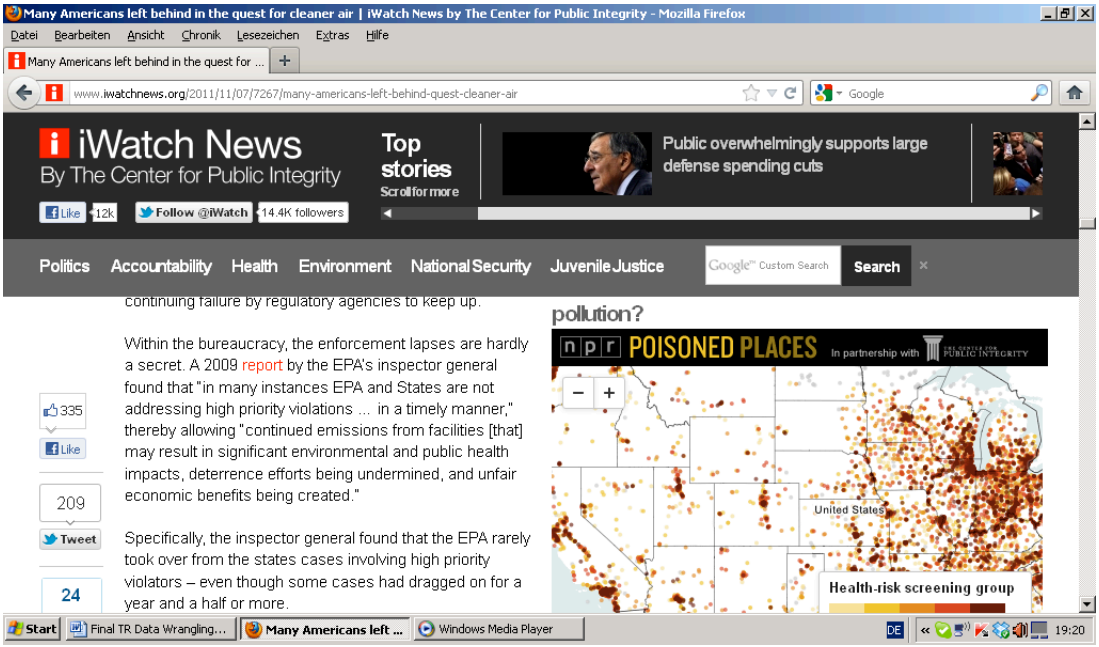
[2] http://ichangemycity.com/

[3] http://tacticaltech.org/studios/projects

The organisation also carries out info-activism camps for campaigners, training them on data-based CSO work and the wrangling of data sets

3. Data-based Journalism

A good example of the combining and analysing of complex data streams for data-driven journalism is the Poisoned Places web-site[4].



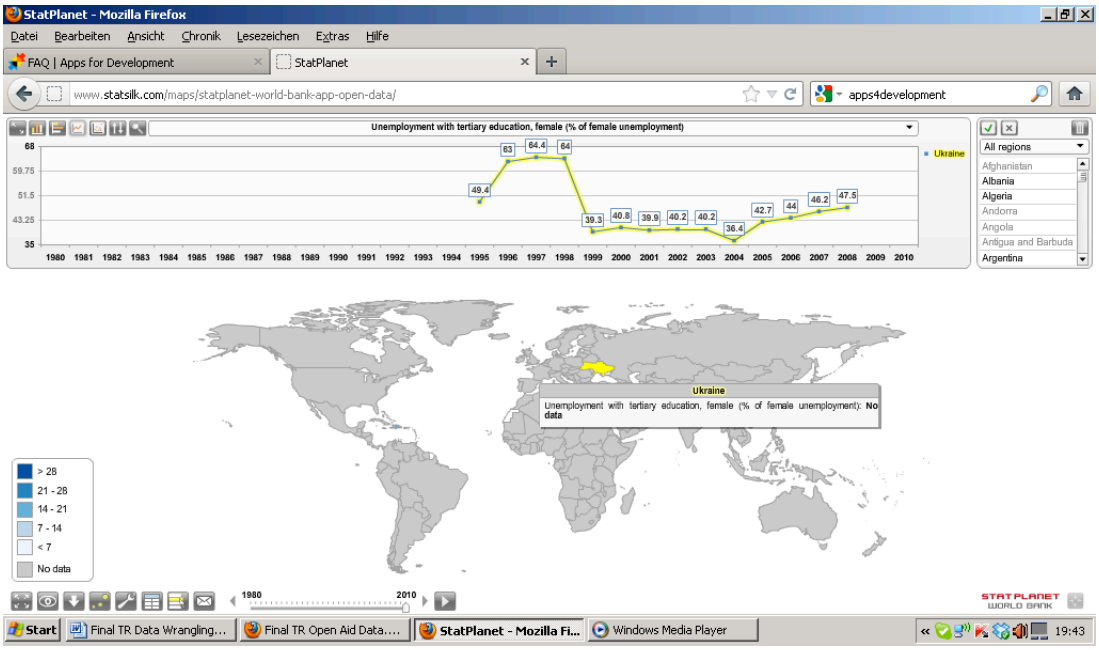
The site combines open PSI such as environmental protection regulations, geodata,

[4] http://www.iwatchnews.org/2011/11/07/7267/many-americans-left-behind-quest-cleaner-air

environmental and health data as well as other statistical indicators which are visualised in various ways, e.g. through interactive maps. In addition to the data, the site features print and video reports on the impacts of air pollution. The site, which is run by the Center for Public Integrity and NPR, was the runner up in the 2012 Data Journalism Awards for Data-Driven Investigations.

4. PSB service delivery

The World Bank Open Data Initiative has been opening its Data Catalogue with approximately 7.000 indicators for re-use and invited contestants to participate in an Apps4Development competition. The winning contribution, StatPlanet[5] combines several thousand indicators which are visualised with maps, graphs and other visualisations.



The opening up of the complex and vast amounts of World Bank data to outside 'wranglers' has helped both the institution itself and other users of the data (e.g. governments, research institutions, media, service providers, and citizens) to make better sense of the data and present it in a more effective manner.

5 Tools and Tasks

Working with data includes tasks such as searching and accessing data, scraping data from web sites, refining, cleaning and parsing data, analysing data, geocoding data and turning data into charts and visualisations.

Currently, there are numerous tools available for data management or wrangling, and new ones are constantly emerging to facilitate this work. These tools include command-line and

[5] http://www.statsilk.com/maps/statplanet-world-bank-open-data

browser-based tools, desktop applications and web-services such as APIs. Some examples of currently available tools include:

Data management tools

Google Fusion Tables is an online database for creating online datasets from different sources, allowing for the visualisation and publishing of the data as maps, timelines and charts.

Google Refine is a powerful software tool for working with "messy" data, cleaning it up, and transforming it from one format into another.

ScraperWiki is an online programming environment for pulling data in from any source, making mash-ups and useful applications out of it.

DataWrangler is an interactive online tool for data cleaning and transformation.

The R Project is a programming language and environment to statistically explore data sets and make graphical representations of the data.

Discussion forums

The website Drawing by Numbers http://drawingbynumbers.org/ is aimed especially at civil society activists and journalists seeking to analyse, manage and visualise data. It provides open source software, "how-to" manuals, toolkits and discussion forums.

Data patterns http://datapatterns.org/ is a web site which includes "tips and tricks for data work," including discussions and manuals for finding, scraping, cleaning and wrangling data.

6 Training courses and curricula for data experts

Several academic institutions, especially in the United Kingdom, offer specialised training courses and curricula on data wrangling and in particular open data

The University of Southampton[6] has been a pioneer in terms of open data and staff from the University are playing in a key role in the setting up of the UK Open Data Institute, which is being supported by the UK Government as part of its Open Government efforts[7]

Also in the UK, Nottingham University has been offering a series of free, one-day "open

[6] http://data.southampton.ac.uk/
[7] http://www.ecs.soton.ac.uk/news/3879

data masterclasses"[8] to a range of possible re-users of PSI, including interested individuals, communities, grassroots organisations, NGOs as well as civil servants and professionals.

The planned School of Data[9] initiative of the Open Knowledge Foundation and Peer 2 Peer University aims specifically at teaching data wrangling skills. Target audiences include interested citizens, journalists, civil society activists, researchers and web developers.

7 Conclusion

With the exponential increase in the amount and quality of data available, data wrangling is becoming increasingly central for individuals and organisations dealing with these information flows. Combining new wrangling techniques with the opening of data potentially allows for new and innovative combinations, analyses, presentations and re-uses of data sets. Given the vastness of PSI data, the variety of forms which it comes in and its centrality to the activities of PSBs, citizens, organisations and corporations, the potential benefits of effective data wrangling tools are immense. However, given the explosive growth in data streams, new and effective forms of data management are not merely a "nice-to-have" side benefit; rather, they are fast becoming essential to keep institutions and processes going.

The combination of PSI re-use with new and effective data wrangling tools has been shown to have benefits for the PSBs, for citizens, CSOs, media and also for the commercial sector. These have been put to use in a number of applications to date to improve service delivery, efficiency, transparency and also highlight serious short comings, such as health and environmental problems.

About the Authors

Daniel Dietrich was born in 1973 in Frankfurt, Germany. His academic work covers political science, computer science and communication science in Frankfurt and Berlin. He worked as Research Associate at Technical University Berlin, Department of Internet and Society until the end of 2011. He has been working for the Open Knowledge Foundation (OKFN), since 2009 and is Chairman of the German Chapter of the Open Knowledge Foundation. He is the Project Coordinator for the OKF Project Open Definition as well as the Coordinator of the Working Group on Open Government Data and the Working Group on Open Data in the EU. He is the co-founder of the Open Data Network, a non-profit advocacy organisation to promote Open Data, Open Government and Transparency in Germany, Europe and beyond. In 2011 he became Editor of the ePSI platform.

Henri Myrttinen was born in Helsinki, Finland, in 1975. He has been working academically and practically on issues related to political and social transformations with various

[8] http://epsiplatform.eu/content/nottingham-university-open-data-masterclass
[9] http://schoolofdata.org/

institutions and organisations for the past 15 years. Geographically, his focus has mostly been in Central and Eastern Europe, Southeast Asia and Sub-Saharan Africa. He received his Ph.D. from the University of KwaZulu-Natal, South Africa, in 2011 and is currently working in Berlin with the Mauerpark Institute.

Copyright information