European Public Sector Information Platform

Topic Report No. 2011/2

Data Journalism Fuelling PSI Re-use

Author: Tom Kronenburg

Published: October 2011

Keywords

PSI Directive, data journalism, data-driven journalism, Open Data

Abstract

PSI is of relevance for journalists because it tells us how our government works. When PSI becomes easier to access and process, it becomes more attractive for journalists to use it as a base for their stories. Open Data is not only easily available but also of impressive quality. Consequently, a new specialty in journalism is surfacing: data journalism. Just as journalism profits from the availability of Open Data, PSI re-users and the Open Data movement can also benefit from the often eye-catching data visualisations and important stories that journalists find in PSI datasets.

Table of Contents

Abstract

Public sector information (PSI) is of relevance for journalists because it tells us how our government works. When PSI becomes easier to access and process, it becomes more attractive for journalists to use it as a base for their stories Open Data is not only easily available but also of impressive quality. Consequently, a new specialty in journalism is surfacing: data journalism. Just as journalism profits from the availability of Open Data, PSI re-users and the Open Data movement can also benefit from the often eye-catching data visualisations and important stories that journalists find in PSI datasets.

Content

In this topic report we aim to explore the relationship between PSI re-use and data journalism. We will first explore the way in which PSI re-use and data journalism are connected, how data journalists work, and what type of PSI-based products are produced by data journalists.

1 What is the relationship between data journalism and Open Data?

Now that more and more data is becoming available for re-use, the Open Data movement is one of the major contributors to the growth of data journalism. As journalists can find value in almost any dataset, the increased availability of Open Data considerably enhances the opportunities to carry out data journalistic projects.

1.1 Data journalists re-use Open Data

Many data journalists use data that is released by governments. For example, journalists who use data on air quality, pollution, crime rates, educational quality and many other subjects are primarily using data that was recorded by public sector bodies (PSBs) or under the responsibility of a PSB. Journalists also use data from trade registers, meteorological offices and many other governmental sources. In the remainder of this Topic Report we will cover a host of examples of government-provided Open Data, highlighting how journalists use government data. Here are five examples of data journalists using Open Data.

1. The Guardian (UK) reported on many British MPs whose expenses were out of line. These reports were based on data obtained from the Parliament.
2. Farmsubsidy.org reports on EU farm subsidies and their appropriation.
3. Many stories on the Iraq and Afghanistan wars were produced based on statistics compiled by the Allied Forces.
4. Journalists throughout Europe compile reports from local police and fire brigade records into databases and subsequently report on the data they have collected.
5. Many news items on environmental quality use data that is collected by meteorological and environmental agencies within governments.

Other than the easy availability, the high quality of Open Data is also a major factor in promoting data journalism. Open Data has a number of characteristics that are very valuable to data journalists, that is, the data:

1. is usually recent and frequently updated;
2. is structured and therefore easy to work with;
3. has a degree of accuracy that no other dataset can ever achieve because it is a product of the PSB releasing it;
4. will usually be updated in the same format, making it much easier to analyse data that has been compiled over time;
5. is formatted in machine-readable formats;
6. is easy to find with the aid of data catalogues;
7. is available at no costs; and
8. is governed by legislation (PSI Directive, FOIA), enabling journalists to use the data without having to seek permission.

Even though any given dataset may not necessarily satisfy all conditions mentioned here, they usually meet a considerable number of these conditions.

1.2 A shared passion for transparency

The Open Data movement has a strong emotional connection to transparency and civic control of government. Both data journalists and Open Data advocates believe monitoring governments and PSBs is part of their task. As such, the emotional connection that data-driven journalists feel towards Open Data is quite strong. In fact, much work is done in both communities that can be described as Open Data activity as well as data journalism. For example, whenever the Open Data community builds services or products specifically tailored towards monitoring government and PSBs, we might very well consider them to be data journalists. Even though they won't feature in a newspaper or journal article, the results of such actions are that citizens who are ordinarily 'out of the loop' obtain a more profound understanding of the processes and decisions made within public institutions. Data journalism is no different from traditional journalism that aims to explain and report on government. One example of Open Data tools that are used by journalists is the host of 'spending' applications that now spawn across Europe. They have already made an impact and resulted in a considerable number of articles and reports.

1.3 Open Data benefits from data journalism

There are a number of benefits that the Open Data community gains from the growing attention to data journalism.

Firstly, the visualizations created by data journalists are often used to prove how astonishing the results of Open Data can be. A good infographic based on PSI beautifully illustrates the point that there is value in datasets that are usually kept hidden.

Secondly, data journalism sometimes leads to the creation of Open Data. The Dutch government is one example of this, as it releases every dataset and document that is requested in FOIA requests on a specific website. This website thus contains a collection of

Open Data documents and datasets for everyone to use.

Open Data can also be created by data journalists. Danish subject and data journalist Tommy Kaas creates openly available databases of police dispatches, fires and other incidents in Denmark. Fishsubsidy.org compiles data from the Member States on EU subsidies to fishing companies into one database that anyone can use. Environmental groups have used this data to create maps of spending and have combined the subsidies data with the fishing quota data. The journalists in these examples take data that is accessible, but not yet re-useable. They then transform the data and offer it to the wider Open Data and data journalism communities.

1.4 Data journalism without PSI and Open Data

Even though we have now seen that a strong bond exists between data journalism and Open Data, we can't say that data journalists rely entirely on governmental Open Data. They can use data from other sources, which cannot be considered to be public sector information.

Firstly, journalists can write stories based on datasets released by commercial organisations. Trends in housing prices and mortgages, car sales, economic indicators and the annual reports of large corporations are all food for the hungry data journalist and most certainly not PSI. Data of this type is usually referred to as 'open corporate data'.

Secondly, journalists have always used confidential and 'leaked' documents as sources for their reports. One of the most recent and striking examples was the usage and release of the Wikileaks cables, which have provided so many leads for journalists. Confidential data however, can never really be 'Open Data' as there is no formal licence under which the data is released.

2 How do data journalists use data?

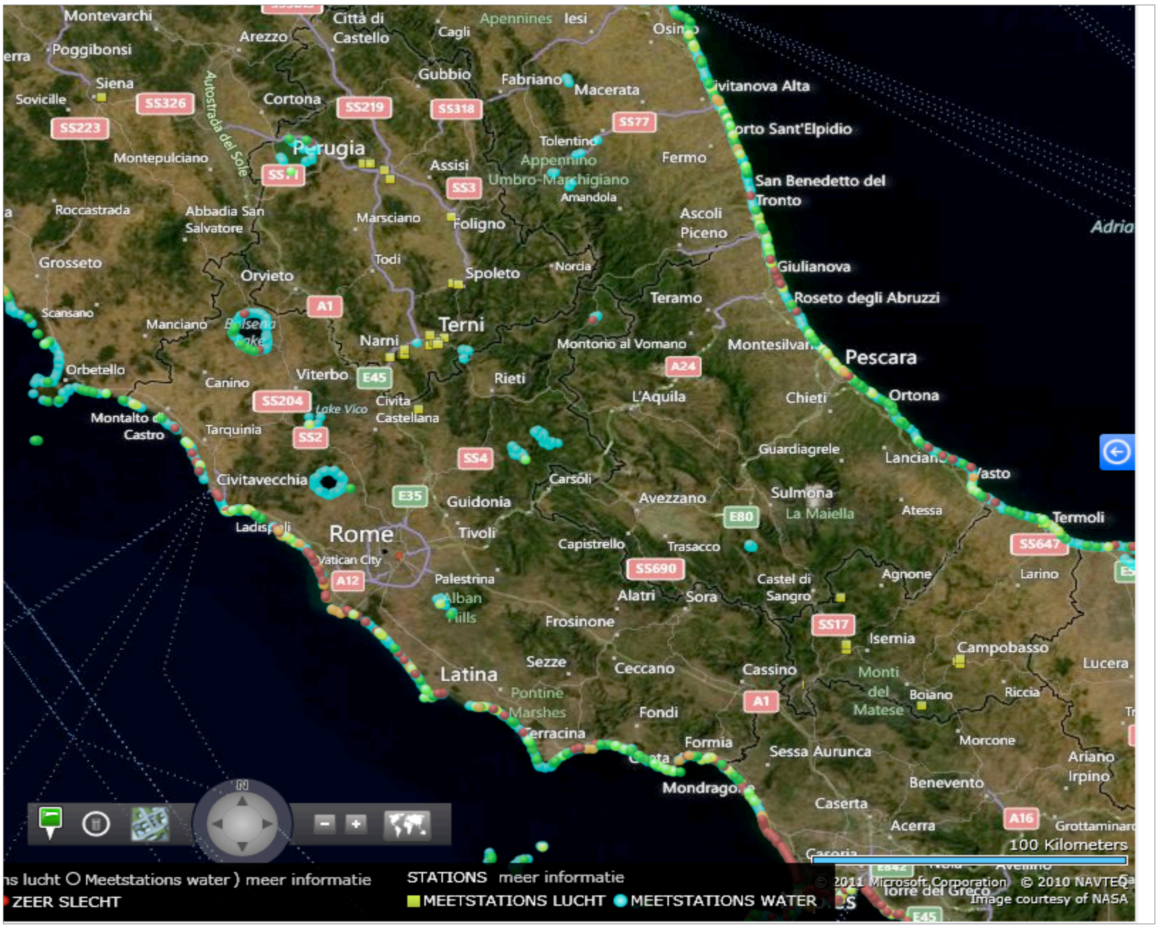
Now that we have established that Open Data has a close relationship with data journalism, we would like to explore the concept of data journalism a little more. What exactly is it that data journalists do with the (open) data they obtain?

There are four basic types of data journalistic activities that use Open Data. All four types can use PSI, and we will provide examples of how journalists used Open Data to write their stories. Data journalists use (open) data:

1. To discover newsworthy facts or stories.
2. To discover trends hidden in large datasets.
3. To compile datasets for further dissemination to the public.
4. To create data visualisations.

2.1 Discovering stories

Recently, the European Environment Agency (the EEA) released data on bath water quality throughout the EU. The dataset and report featured both quality of water around beaches and in rivers and ponds. The screenshot below shows the tool through which they present the data, and which is used by the EEA for interaction with beach goers. Green and blue dots represent good water quality, red and orange dots bad quality.



Water quality in the area around Rome, Italy.

The release of the bathing water report and dataset[1] spurred an avalanche of newspaper and online articles by local publicists, each focussing on their own local stretch of swimming water (e.g., these articles on Cypriot bathing water).[2] The questions journalists typically answer in their articles about water quality are:

1. Has the quality of swimming water improved or deteriorated?
2. How do our beaches compare to those of our neighbours?
3. Do we see the result of (local) policy reflected in the data provided by the EU?

[1] http://www.eea.europa.eu/data-and-maps/data/bathing-water-directive-status-of-bathing-water-3

[2] http://www.argophilia.com/news/eu-bathing-water-quality/22965/ and http://www.cyprusandmore.com/2011/06/cyprus-has-cleanest-bathing-water-in-europe/)

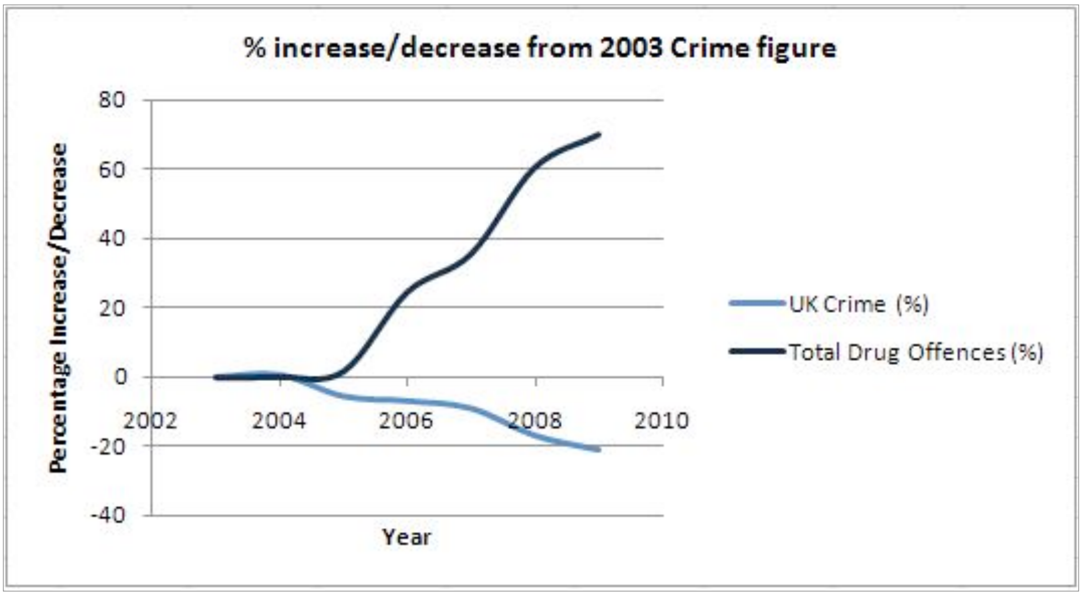
We see that every journalist can discover his or her own local story for the (hyper)local news outlet. The main journalistic process in this example is therefore the discovery of individual news items from a large set of data. Many data journalists can use data in this way, as it is one of the easiest ways of working with data.

Interestingly enough, the anecdotal evidence discovered from the analysis of large datasets is usually quite strong. This stems from the fact that the data journalist can choose the 'best' cases that showcase the problem in its most acute form.

2.2 Discovering trends

Discovering anecdotal evidence, as described above, is not the only purpose for which datasets are very valuable. They can also be used for discovering trends and discerning wider patterns. These trends would not be visible without the data, and they are usually discoverd by creating graphs from datasets or doing some form of statistical analysis.

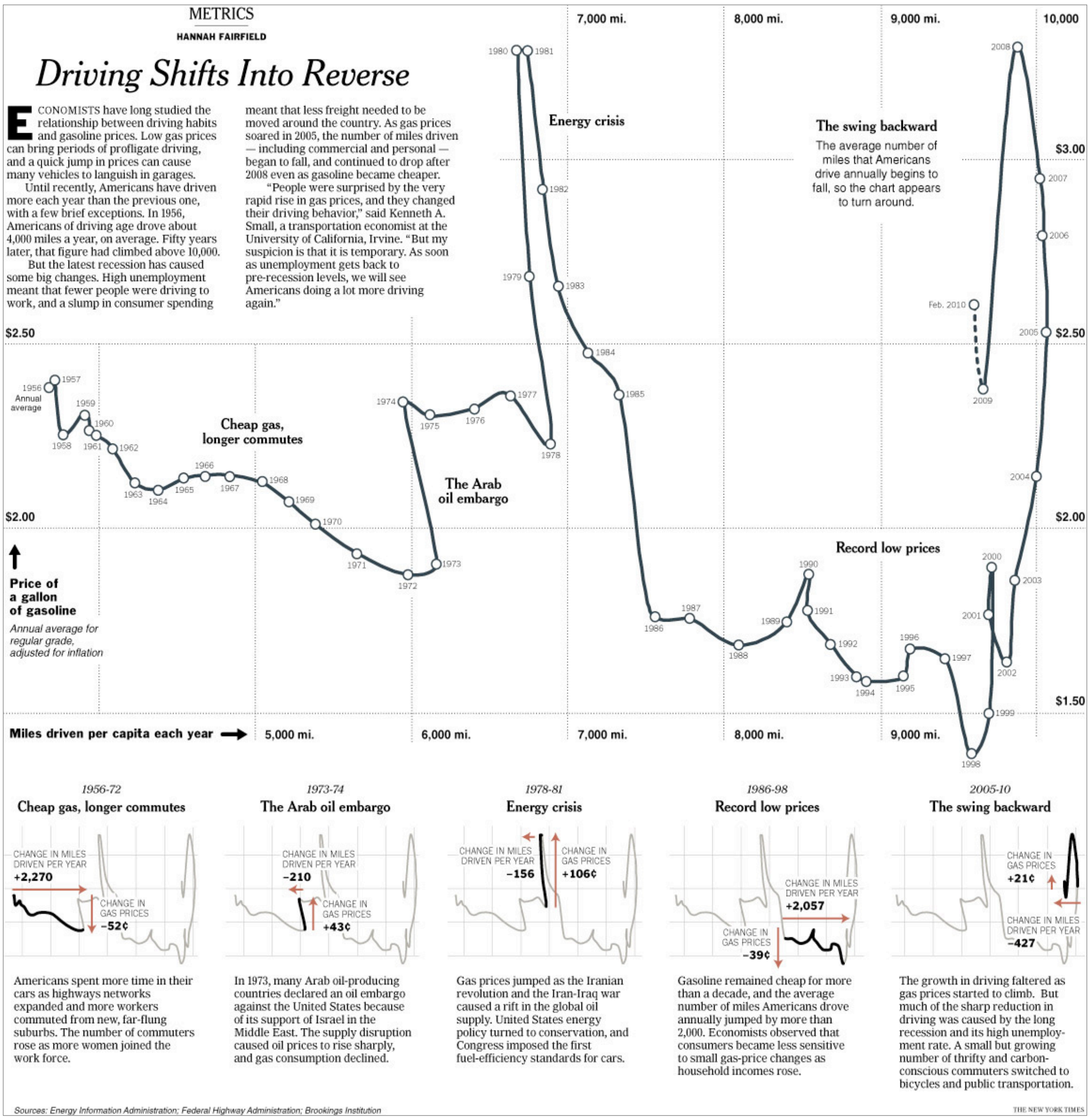
Interesting examples are the war logs of both the Iraq and Afghanistan wars, and the way Hans Rosling uses data to analyze what's happening to the world's population in this TED video[3]. In the bathing water dataset above, journalists looked at the local trends of water quality over time. Spending data also lends itself quite happily toward trend analysis, as it easily shows how spending changes over time. Data journalist Julia Greenway published a very interesting article on crime statistics. She noticed in crime statistics, taken from data.gov.uk, that the total amount of crime was steadily falling although drug-related crime was on the rise. In the article she shows how both trends compare, using a simple graph.



Another excellent example of showing the trends found in data is this visualisation by the New York Times. The journalist clearly shows the influence of a number of historic events

[3] http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

on the gas prices paid by American citizens at the pump. She shows that the mean driving distances in the first decade of the 21st century only marginally decrease, even when the price of gas changes enormously. The visualisation is very attractive, but more important is the discovery and illustration of this very powerful finding. The data for both gas prices and mean driving distances were obtained from the Energy Information Administration, the Federal Highway Administration and the Brookings Institution (all USA).



METRICS

HANNAH FAIRFIELD

Driving Shifts Into Reverse

ECONOMISTS have long studied the relationship between driving habits and gasoline prices. Low gas prices can bring periods of profligate driving, and a quick jump in prices can cause many vehicles to languish in garages.

Until recently, Americans have driven more each year than the previous one, with a few brief exceptions. In 1956, Americans of driving age drove about 4,000 miles a year, on average. Fifty years later, that figure had climbed above 10,000.

But the latest recession has caused some big changes. High unemployment meant that fewer people were driving to work, and a slump in consumer spending meant that less freight needed to be moved around the country. As gas prices soared in 2005, the number of miles driven — including commercial and personal — began to fall, and continued to drop after 2008 even as gasoline became cheaper.

"People were surprised by the very rapid rise in gas prices, and they changed their driving behavior," said Kenneth A. Small, a transportation economist at the University of California, Irvine. "But my suspicion is that it is temporary. As soon as unemployment gets back to pre-recession levels, we will see Americans doing a lot more driving again."

The swing backward

The average number of miles that Americans drive annually begins to fall, so the chart appears to turn around.

1956-72
Cheap gas, longer commutes

Americans spent more time in their cars as highways networks expanded and more workers commuted from new, far-flung suburbs. The number of commuters rose as more women joined the work force.

1973-74
The Arab oil embargo

In 1973, many Arab oil-producing countries declared an oil embargo against the United States because of its support of Israel in the Middle East. The supply disruption caused oil prices to rise sharply, and gas consumption declined.

1978-81
Energy crisis

Gas prices jumped as the Iranian revolution and the Iran-Iraq war caused a rift in the global oil supply. United States energy policy turned to conservation, and Congress imposed the first fuel-efficiency standards for cars.

1986-98
Record low prices

Gasoline remained cheap for more than a decade, and the average number of miles Americans drove annually jumped by more than 2,000. Economists observed that consumers became less sensitive to small gas-price changes as household incomes rose.

2005-10
The swing backward

The growth in driving faltered as gas prices started to climb. But much of the sharp reduction in driving was caused by the long recession and its high unemployment rate. A small but growing number of thrifty and carbon-conscious commuters switched to bicycles and public transportation.

Sources: Energy Information Administration; Federal Highway Administration; Brookings Institution

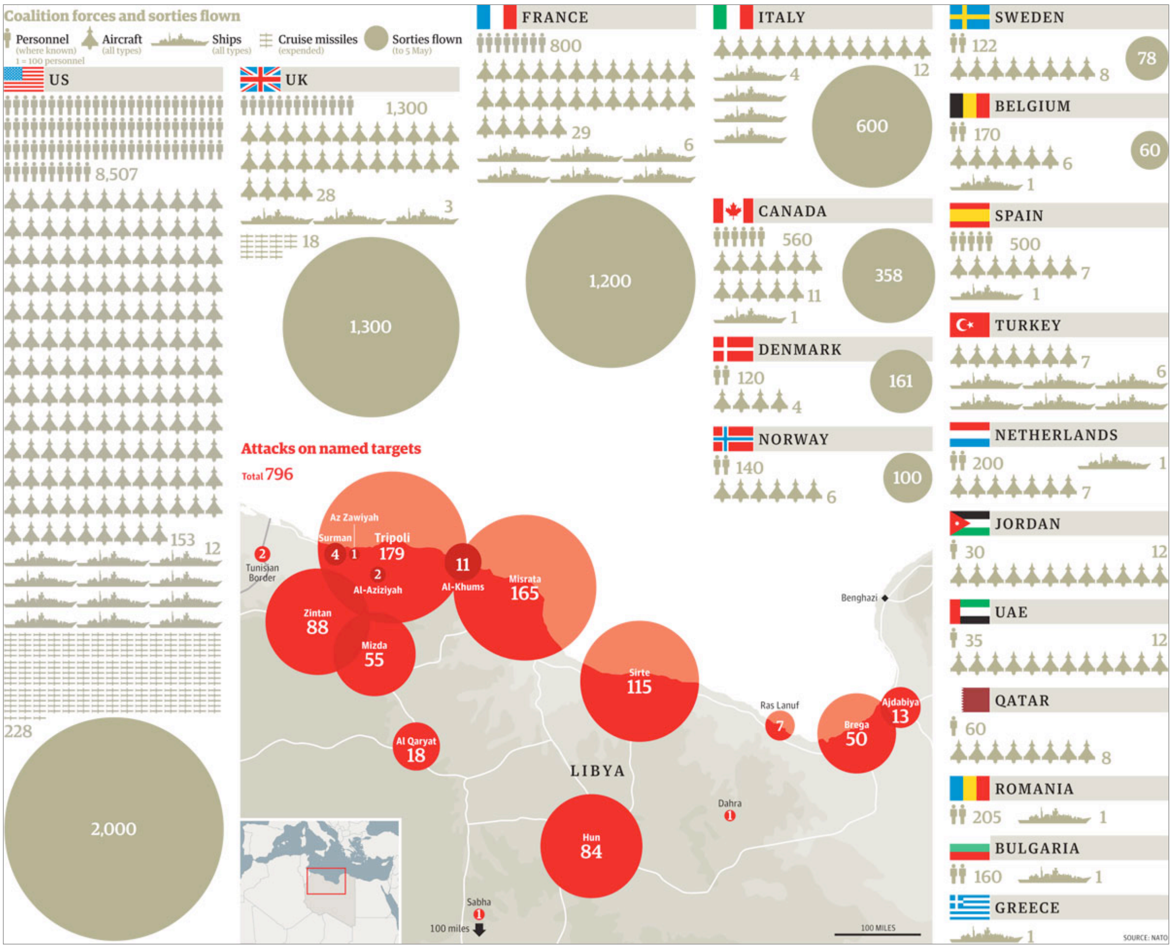
THE NEW YORK TIMES

New York Times article on driving distances and gas prices by Hannah Fairfield.

Discovering trends usually means analysing data from a longer time period. This means that the journalist needs to obtain and analyse more data. Finding trends in data is closely related to the concept of precision journalism that brings an almost academic rigour to journalism. Also, to aid in this task, it is important that datasets are released in similar formats over a longer period of time. This quality is most often found in Open Data, not in other, less formally structured data. Annual reports of both PSBs and commercial enterprises are especially notorious for changing context and format of data over time.

2.3 Visualizing data

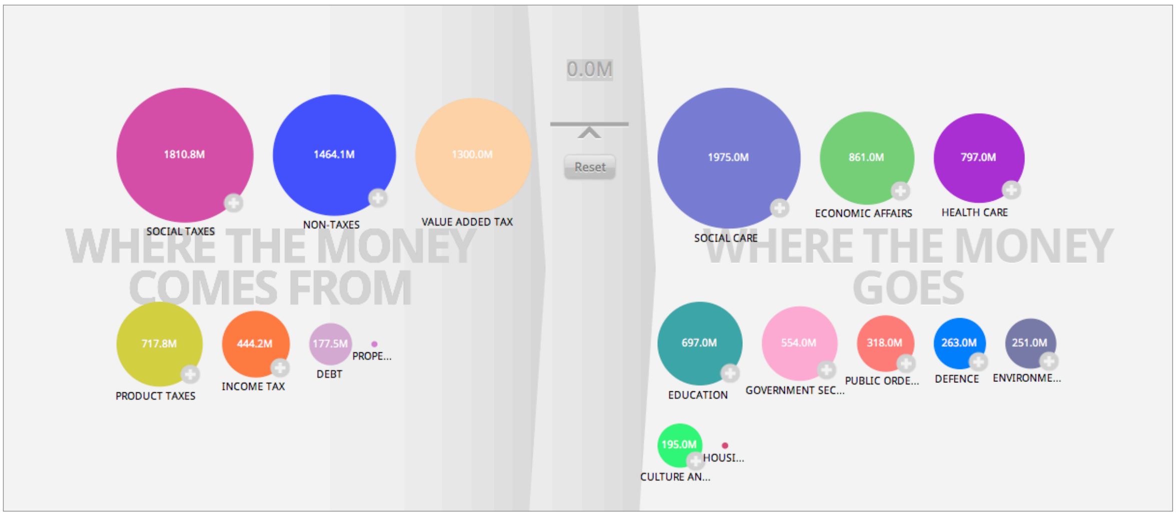
The age-old saying that ‘a picture is worth a thousand words’ still has validity in our world. Making sense of big sets of data is inherently easier when they are presented in well-chosen visual formats. Below, a visualisation from the Guardian on the composition of the Coalition troops currently fighting in Libya against the troops of Col. Qaddafi, under the flag of Operation ‘Unified Protector’, clarifies a lot.



NATO operations within ‘Unified Protector’.

Writing up a story so as to include all of the information contained in this visualisation would probably result in a dull piece and a substantial loss of information, especially the relative effort of all nations involved. The form in which the data is presented has, to a large extent, determined the attractiveness of this story, making it more appealing to a wider audience than it would have been were it presented as written text.

Data journalists also use PSI for news visualisation. The spending application http://meieraha.eu/ shows how the Estonian government budget is allocated.

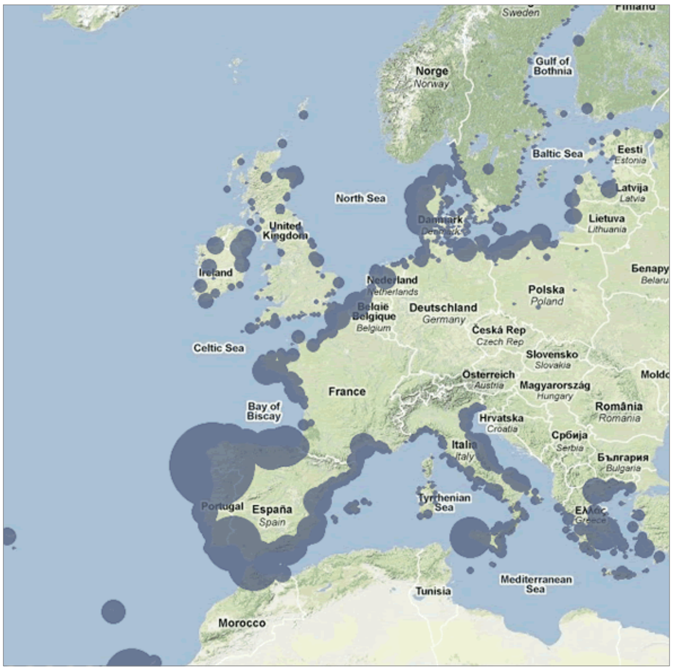


Estonian Government Budgets: Income vs. Spending. Clicking the 'Bubbles' will open more detailed categorizations.

2.4 Disseminating data to the public

While data visualisation or dataviz is all about compiling as much information as possible into a compelling and informative graphic, another branch of data journalism focuses on compiling new sets of data and making them public. A well-known example of this practice is the publishing of the expense accounts of several Members of Parliament (MPs) in full by the Guardian, hoping the public would aid them in their search for fraudulent behaviour by MPs. The Guardian obtained the original data from Parliament and reworked the documents, enabling citizens to easily search the data and report on their findings. Although the data was available earlier (on paper, in mixed formats), the effort of exploring this data to find all the relevant information was too much.

The journalists behind Farmsubsidy.org and Fishsubsidy.org use publicly available data on EU farm and fish subsidies (published by the Member States) and import this into databases. The databases are made available to anyone who wants to further investigate the subsidies. The effort also created the opportunity to cross reference subsidies from different Member States.



Fish subsidy data from all Member States shows where EU fishing subsidies go.

In light of this trend of compiling datasets, some people now consider the task of data journalists to be “supplying databases with raw material — articles, photos and other content — by using medium-agnostic publishing systems and then making it available for different devices."[4] In 2006, Holovaty wrote the seminal piece ‘A Fundamental Way Newspaper Sites Need to Change’[5]. In this article, Holovaty wrote that most material collected by journalists is "structured information: the type of information that can be sliced-and-diced, in an automated fashion, by computers". He considered it a journalist’s task to discover the structure and to make the underlying data available in such a way that readers can analyse and explore the information themselves. The end result of the journalistic process is, in Holovaty’s opinion, a database of structured information. This thinking is in line with the approach of the linked data movement, another movement with strong ties to the Open Data community.

3 Where to find European data journalists and how to join their ranks?

If you have a personal interest in data journalism, we urge you to just start exploring. It can be really easy! First, find some datasets. The national data catalogues are a great starting point. Load the datasets into a spreadsheet program and off you go. You can start exploring the data and combining it, finding trends or creating visuals. Even a simple graph depicting a trend can be an interesting and newsworthy fact. If this small manual is not enough for you, we encourage you to look around on the web. We give below a brief list of interesting websites, organisations, books and articles that can be of assistance.

3.1 Organisations

The European Journalism Centre (http://www.ejc.net/ejc/) often has stories, events and other resources on data journalism. They also have a data journalism community on http://community.ejc.net/group/datadrivenjournalism.

The Pascal de Croo Fund in Brussels (http://www.fondspascaldecroos.org/) is heavily involved with the EU data journalism scene, often organizing events and courses. http://www.wobbing.eu is its main hub for data-driven journalistic activities.

The farm subsidy (http://farmsubsidy.org/) and the fish subsidy (http://fishsubsidy.org/) projects involve a number of data journalists from around the world, but mostly from the EU.

The Open Knowledge Foundation (http://www.okfn.org) has a data journalism working group, manages a mailing list on data journalism and often speaks on data journalism topics.

[4] http://131.193.153.231/www/issues/issue7_8/loosen/

[5] http://www.holovaty.com/writing/fundamental-change/

The Organized Crime and Corruption Reporting Project ((OCCRP) http://www.reportingproject.net/) is a joint program of the Center for Investigative Reporting in Sarajevo, the Romanian Center for Investigative Journalism, the Bulgarian Investigative Journalism Center, the Center for Investigative Reporting in Serbia, the Caucasus Media Investigation Center, Novaya Gazeta, HETQ in Armenia and a network of investigative journalists and media in Montenegro, Albania, Moldova, Ukraine, Macedonia and Georgia. A considerable amount of data journalism activity is taking place within the OCCRP.

The Global Investigative Journalism Network (http://www.globalinvestigativejournalism.org) is involved in a number of data journalism projects.

The Association of European Journalists – Bulgaria (http://www.aej-bulgaria.org/) is involved in a number of data journalism projects.

The Guardian Data Blog (http://www.guardian.co.uk/news/datablog) remains one of the prime sources for anyone interested in data journalism. The Guardian has a sizable section of data journalists and visualisation artists and sets the standard in the world of data journalism.

Other large EU news agencies that have an interest in data journalism are Aftenposten (Norway), Financial Times (UK) and Die Zeit (Germany).

3.2 Online resources

http://datadrivenjournalism.net/ Report from the 2010 conference on data journalism in Amsterdam. An interesting conference report is available at: http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf

http://datajournalism.stanford.edu/ A very entertaining video project on data journalism by Stanford scholar and long-time journalist Geoff McGhee.

http://nrk.no/maktbasen/index.php?lcc=shell Interesting project showing relationships between Norwegian companies and politicians.

http://www.guardian.co.uk/news/datablog The data blog of the British newspaper The Guardian.

http://blog.zeit.de/open-data/ The data blog of German newspaper Die Zeit.

http://mps-expenses.guardian.co.uk/ An innovative crowdsourcing application allowing users to check expense documents, adding indications on whether the documents should be investigated further or not.

http://flowingdata.com/ A site showcasing many beautiful visualisations. Also check: http://www.informationisbeautiful.net/ for more good visuals.

Two TED videos on data visualisation: Hans Rosling (2006) and David McCandless (2010) http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

http://www.informationisbeautiful.net/visualizations/the-billion-dollar-gram/ Information is Beautiful: 'The Billion Dollar Gram' — billions of costs in a treemap, showing the relationship between issues and spending.

http://www.mediastorm.com/publication/never-coming-home Mediastorm: 'Never Coming Home', turns the data points of casualties in the US Army into personal stories.

http://www.nytimes.com/interactive/world/war-logs.html The New York Times: 'The War Logs', a structured and extended overview on the complex issue of the wars in Iraq and Afghanistan.

http://www.wobbing.eu details FOIA requests across the EU.

3.3 Introductory Articles

(from http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf)

A short list of must-read articles on data-driven journalism. The goal here is to guide newcomers to some insightful pieces, without trying to be complete. Feel free to send us recommendations for additional input, best done by posting your comments on the data-driven journalism group at the EJC website.

Adrian Holovaty, 'A fundamental way newspaper sites need to change', Holovaty.com, Sept. 6, 2006, http://www.holovaty.com/writing/fundamentalchange/

Stijn Debrouwere, 'We're in the information business', April 2010, stdout.be, http://stdout.be/2010/we-are-in-the-information-business/

Eric Ulken, 'Building the data-desk: lessons from the L.A. Times', The Online Journalism Review, Nov. 21, 2008, http://www.ojr.org/ojr/people/eulken/200811/1581/

'Journalism Needs Data in the 21st Century', ReadWriteWeb, Aug 5, 2009,

http://www.readwriteweb.com/archives/journalism_needs_data_in_21st_century.php

Rich Gordon, 'What Will Journalist-Programmers Do?', MediaShift Idealab, Nov. 18,

2007, http://www.pbs.org/idealab/2007/11/what-will-journalist--programmersdo005.html

Rich Gordon: 'Data as journalism, journalism as data', Readership Institute, Nov. 14,

2007, http://getsmart.readership.org/2007/11/data-as-journalism-journalism-asdata.html

3.4 Data Books

(from http://mediapusher.eu/datadrivenjournalism/pdf/ddj_paper_final.pdf)

Nick Davis, *Flat Earth News,* Random House UK, 2008.

Criticism of 'churnalism' in newspapers — some very revealing back-stories on how the truth is sometimes distorted by journalists. The most revealing story might be the crusade against heroin, which may have unintentionally helped to create a drug market. If the journalists had known the data, would they have argued differently? ISBN 9780099512688

Robert McKee, *Story*, HarperCollins Publishers, 1997.

An indispensable handbook for screen and storywriters. If multimedia journalism is to grow we need more knowledge on how to do it right. ISBN 9780060391683

Dan Roam, *The Back of the Napkin*, Penguin Group, USA, 2008

Helps you to develop new skills in visualizing (by hand), but can be a good creative source for the future. ISBN 9781591843061.

Ian Ayres, *Super Crunchers, Why Thinking-by-Numbers Is the New Way to Be Smart*, Random House Publishing Group, 2008.

Sums up and describes techniques for data mining and introduces examples of how big data can help to make predictions for the future. ISBN 9780553384734

Malcolm Gladwell, *Outliers. The Story of Success*, Little Brown & Company, 2008.

Essentially most of these stories are data-driven stories; the findings are sometimes amazing and fun to read. For example, the 10,000-hour rule that may help to understand the successful careers of programmers and musicians such as the Beatles. ISBN 9780316017930

About the Author

Tom Kronenburg is a consultant with Zenc B.V. based in the Netherlands. He specialises in information as a solution to societal problems. Tom is one of the curators of the EPSI Platform website and travels throughout the European Union to connect PSI holders and re-users, citizens and governments.

Copyright information

This document and all material therein have been compiled with great care. However, the author, editor and/or publisher and/or any party within the European PSI Platform or its predecessor projects the ePSIplus Network project or ePSINet consortium cannot be held liable in any way for the consequences of using the content of this document and/or any material referenced therein. This report has been published under the auspices of the European Public Sector Information Platform.

The report may be reproduced providing acknowledgement is made to the European Public Sector Information (PSI) Platform. The European Public Sector Information (PSI) Platform is funded under the European Commission eContentplus programme.