



D2.3 BEST PRACTICE VISUALISATION, DASHBOARD AND KEY FIGURES REPORT

PROJECT

Acronym: **OpenDataMonitor**
Title: Best practice visualisation, dashboard and key figures report
Coordinator: SYNYO GmbH

Reference: **611988**
Type: Collaborative project
Programme: FP7-ICT

Start: November 2013
Duration: 24 months

Website: <http://project.opendatamonitor.eu>
E-Mail: office@opendatamonitor.eu

Consortium: **SYNYO GmbH**, Research & Development Department, Austria, (SYNYO)
Open Data Institute, Research Department, UK, (ODI)
Athena Research and Innovation Center, IMIS, Greece, (ATHENA)
University of Southampton, Web and Internet Science Group, UK, (SOTON)
Potsdam eGovernment Competence Center, Research Department, Germany, (IFG.CC)
City of Munich, Department of Labor and Economic Development, Germany, (MUNICH)
Entidad Publica Empresarial Red.es, Shared Service Department, Spain, (RED.ES)

DELIVERABLE

Number:	D2.3
Title:	Best practice visualisation, dashboard and key figures report
Lead beneficiary:	ODI
Work package:	WP2: Research studies and stakeholder analysis
Dissemination level:	Public (PU)
Nature:	Public (R)
Due date:	July 31, 2014
Submission date:	July 31, 2014
Authors:	Ulrich Atz , ODI Tom Heath , ODI Michael Heil , SYNYO Jack Hardinges , ODI Jamie Fawcett , ODI
Contributors:	Yunjia Lee , SOTON Peter Leitner , SYNYO Amanda Smith , ODI
Reviewers:	Ejona Sauli , SYNYO Bernhard Jaeger , SYNYO

Acknowledgement: The OpenDataMonitor project is co-funded by the European Commission under the Seventh Framework Programme (FP7 2007-2013) under grant agreement number 611988.

Disclaimer: The content of this publication is the sole responsibility of the authors, and in no way represents the view of the European Commission or its services.

SUMMARY

Various initiatives have emerged attempting to benchmark different aspects of the open data ecosystem, each placing a different emphasis or adopting different methodologies. The OpenDataMonitor project takes a particular perspective, focusing on automated assessment of open data deployment across Europe, as determined by analysis of the (meta)data available in open data catalogues.

This report presents the suite of metrics that the OpenDataMonitor platform will compute based on harvested metadata. These metrics are classified at the following levels of (decreasing) granularity: 1) measures over the aggregate; 2) per-geography measures; 3) per-catalogue measures; and 4) per-dataset measures.

To aid in gaining insights from these metrics, the report examines a range of visualisation techniques that may be employed to present them in graphical form. These are mapped to each metric to guide implementation of the OpenDataMonitor platform, coupled with a survey of software visualisation libraries that may be used in the project.

Recognising the need to present metrics and visualisations in coherent groups rather than in isolation, the report concludes with a review of information dashboard techniques and best practices, including recommendations of how these may be adopted in generating dashboards for users of the OpenDataMonitor platform.

TABLE OF CONTENTS

1	Introduction	7
2	Monitoring the open data landscape	8
2.1	Existing open data monitoring approaches.....	8
2.2	Gold-standard characteristics of open data	11
3	Metrics and key figures for OpenDataMonitor	14
3.1	Measures over the aggregate	15
3.2	Per-geography measures.....	19
3.3	Per-catalogue measures.....	21
3.4	Per-dataset measures.....	28
3.5	Summary	29
4	Visualisation techniques for monitoring open data	30
4.1	Introduction to data visualisations.....	30
4.2	Guidance and principles in applying visualisation techniques	31
4.3	An example of applying the principles of good visualisation design	34
4.4	Visualisations of the open data ecosystem	36
4.5	Selection of visualisation techniques	55
4.6	Mapping of metrics to visualisation techniques	57
4.7	Software Libraries for Visualisation.....	63
5	Creating a dashboard	69
5.1	Introduction to dashboards.....	69
5.2	Recommended techniques and examples	70
5.3	Dashboarding requirements of OpenDataMonitor.....	75
6	References.....	76

LIST OF FIGURES

Figure 1. Mapping classes of open data metric to preliminary dimensions of availability and measurability.....	14
Figure 2. William Playfair’s visualisation of wheat prices.	30
Figure 3. Internet use among adult men and women in Scotland.....	34
Figure 4. Internet use among adult men and women in Scotland (Redesigned from ODI)	35
Figure 5. ePSI scoreboard.....	37
Figure 6. Graduated symbol map.....	38
Figure 7. Cartogram: regional geography of peer-to-peer lending in the UK.....	38
Figure 8. PSI overall score	40
Figure 9. Redesigned bar chart with emphasis on the Netherlands (Source: ODI).....	41
Figure 10. Example bar chart from the ENGAGE project	42
Figure 11. Example of a stacked bar chart	44
Figure 12. An example of Maturity progression. Veljković et al. (2014).....	46
Figure 13. Example histogram.....	47
Figure 14. Example pie chart (Source: http://www.engagedata.eu/opendatasites).....	48
Figure 15. Example network visualisation (Source: https://data.cityofnewyork.us/viz)	49
Figure 16. Example heat map (Source: https://index.okfn.org/country)	50
Figure 17. Example tabular heat map with continuous scales.....	50
Figure 18. Variations of the heat map.....	51
Figure 19. Example scatter plot (Source: http://www.gapminder.org/)	52
Figure 20. Example spark line.....	53
Figure 21. Example of small multiples.....	53
Figure 22. Example composite graph	54
Figure 23. Example bespoke infographic	55
Figure 24. Chart Chooser.....	56
Figure 25. Google Trend analysis of the most common charting libraries of the past 10 years.....	66
Figure 26. An excerpt of the possibilities that D3.js provides as shown on the solution’s home page	66
Figure 27. An excerpt of the chart types	67
Figure 28. Interactivity features	67
Figure 29. Combination of a choropleth map and a line chart using Highcharts.....	67
Figure 30. A highly flexible and interactive stacked area chart rendered by NVD3	68
Figure 31. A time series chart where it can be.....	68
Figure 32. Visualisation of students’ performance in a school class.....	71
Figure 33. Open Data Institute Company Dashboard	72
Figure 34. (Source: https://index.okfn.org/).....	73
Figure 35. London City Dashboard	73
Figure 36. Eurostat regional statistics explorer.....	74

LIST OF TABLES

Table 1. Eight principles of open government data	11
Table 2. Feasibility of measuring the eight principles	12
Table 3. Counts / averages / longitudinal	15
Table 4. Rankings	17
Table 5. Templates for summary statistics for aggregated lower-level metrics	18
Table 6. Catalogue statistics across geographical regions	19
Table 7. Time-profile by country	20
Table 8. Data volumes	21
Table 9. Data duplication/uniqueness	22
Table 10. Housekeeping	23
Table 11. Formats and machine-readability	23
Table 12. Licenses	25
Table 13. Release frequencies and timeliness	26
Table 14. Prominence, engagement and usability	27
Table 15. Data and metadata volume, quality and usability	28
Table 16. Pitfalls of data visualisations and what to avoid	32
Table 17. The Friendly Data Graphic, Tufte (1983)	36
Table 18. Variations of the cartogram	39
Table 19. Variations of the bar chart	42
Table 20. Variations of the stacked bar chart	45
Table 21. Variations of the line chart	46
Table 22. Variations of the pie chart	48
Table 23. Overview of ODM's main visualisation techniques	56
Table 24. Overview of all JavaScript visualization libraries and frameworks that are open source	64

1 Introduction

Recent growth in deployment of open data has presented myriad opportunities for new economic activity, for greater efficiency and transparency, and for increased citizen engagement. However, with growth comes the challenge of understanding an increasingly diverse and complex landscape, such as the availability of open data.

OpenDataMonitor (ODM) will reveal where, when, how, and by whom open data is being deployed across Europe. This will be achieved by developing and delivering an analysis and visualisation platform that harvests metadata from local, regional and national data catalogues and provides insights into open data availability and publishing patterns. The platform will help developers, entrepreneurs, civil society, policy makers and enthusiasts to understand how the open data ecosystem is evolving and to discover sources of open data that are appropriate to their needs.

The value of these analytics to the platform's target audiences will be strongly dependent on the quality of the visualisations and dashboards used to present the results of the analytics, as well as on the suitability of the underlying metrics themselves.

The primary aims of this report are to:

- present a set of metrics relevant to understanding trends in open data deployment – these will form the basis for the analytics provided in the platform;
- detail the visualisation techniques that may be used to present these metrics, and their underlying trends, in visual form;
- recommend how these visualisations may be combined into dashboards for presentation to end users.

Each element of the report will, where appropriate, be supported by relevant literature and related work that has informed the guidance provided here.

2 Monitoring the open data landscape

OpenDataMonitor is not the first attempt to systematically monitor the open data landscape. In this section we will discuss a few examples of previous work, commenting on their methodology and pointing out the specific focus chosen by the researchers. This overview shows how there is a wide variety of possible approaches. In general, we can imagine two extremes: a “story-driven” case study approach and an automatic quantitative analytics report. Given the scope of OpenDataMonitor we will focus on the examples that are closer to the latter, hence, quantitative in nature. Additional background on existing studies is provided in Deliverable 2.2.

2.1 Existing open data monitoring approaches

2.1.1 Open Data Barometer

The Open Data Barometer measures the distribution and impact of open government data policies and practices in 77 countries around the world. Run in 2013 as a joint project between the Open Data Institute and the World Wide Web Foundation, the Barometer uses multidimensional analysis to score countries’ overall progress in realising the potential benefits of open data as well as across a range of categories (such as education or environment). This is achieved through exploration of each country’s structural readiness to benefit from open data, the extent to which key government datasets are published and the measurable political, social and economic impacts open government data has had.

Key points relevant to OpenDataMonitor

- Uses “peer-reviewed expert **surveys**” and “secondary data sources” - therefore not automated
- Geography - limited to 77 countries but quite **high coverage of EU**
- Scope is **limited to National-level analysis** - not particular catalogues
- Slight **focus on the release of govt data** but there are provisions for the USE of govt data by business and citizens
- Limited visual **data exploration tool** providing a spider-diagram of dimension scores and different sized dots for each of the categories

2.1.2 Open Data Index

The Open Data Index provides an annual report on the global state of governmental open data release. The Index platform is compiled by the Open Knowledge Foundation and uses the data collected by the Open Data Census. It provides a global score comparison by aggregating 10 open data categories for the 70 countries it covers.

With regard to visualisation, the Open Data Index presents a simplified graphic report based upon expert data assessment, although this platform also allows for further user data exploration and user contribution to the Open Data Census to alleviate out-dated data and/or propose information to be included in later editions.

Key points relevant to OpenDataMonitor

- Allows for **quantitative score comparison** but the **scoring methodology used is non-automated** and **conducted by experts**
- **Global focus** adopted, in line with that of the Open Data Barometer
- Reports on **government data only**
- The platform **considers 12 metrics** (9 binary and 3 qualitative) and **weights these differently** to calculate each country's score
- The raw **data is made available in CSV and JSON formats** and the platform uses **open source code**

2.1.3 Open Data Compass

The Open Data Compass, produced by a company called Arachnys, measures the availability of open legal and structural data on companies worldwide. Individual countries are scored, and subsequently ranked, on how comprehensive, online, freely available and conveniently searchable their corporate and litigation records are, as well as the size of their news media. The tool is primarily aimed at professionals to gain insight into foreign markets rather than interrogate the quality of open data per se, however it does to some extent suffice in this regard.

Key points relevant to OpenDataMonitor

- Scores for Corporate and Litigation are determined **qualitatively**, and News also not **non-automated**
- Geography - **215 countries**
- **Restricted to corporate sector** - specifically **limited to legal** issues surrounding registration and litigation of companies (as opposed to their own publication of datasets)
- Limited to **national level** analysis
- Relatively good **data exploration tool**

2.1.4 Open Data 500

The Open Data 500 study is conducted by GovLab and the platform presents a descriptive summary of the open data usage of US companies. The study uses a combination of outreach campaigns, expert advice and research to identify 500 companies to be included and documents their use of open government data. The broader goal of the programme is to assess the economic value of open government data, catalyse the development of open data businesses, and to enhance dialogue between government and business regarding open data usage.

The Open Data 500 platform adopts a data exploration tool to allow for user exploration using state map, data category and federal agency filters. In addition, the primary information, the open data flow from federal agencies to US companies, is presented in interactive compass form.

Key points relevant to OpenDataMonitor

- The comparison platform collects **descriptive-based, qualitative data** on the **private sector usage of government data only**
- It adopts **non-automated, self-certified surveying** and incorporates **7 key metrics**
- Platform has a **US focus only**

- **Raw data is made available in CSV and JSON formats** although the platform uses **unspecified and unavailable code**

2.1.5 Metadata Census

The Metadata Census assesses, scores and ranks the quality of metadata in open government data repositories around the world. It does this by judging the metadata entries in each repository on a number of quality criteria including accuracy, availability and completeness - including importantly whether entries lead to the correct datasets. It carries out this function in order to measure the repositories effectiveness in supporting the tasks of finding, identifying, selecting and obtaining datasets.

Key points relevant to OpenDataMonitor

- Highly **relevant** to Open Data Monitor
- **Automated metadata harvester** - uses snapshots - updated but not continuously
- Currently only **government data** - only CKAN repositories - relatively **good worldwide** coverage however
- **8 metrics** - completeness, weighted completeness, accuracy, richness of information, readability, availability, misspelling, openness
- However - by its own admission tends to **overvalue issues** with metadata and is limited in implementing a range of quality metrics
- Analysis at **repository level**
- **Data exploration tool**

2.1.6 Open Data Certificates

The Open Data Certificates is a platform operated by the Open Data Institute that seeks to provide detailed information on open datasets and thus increase the level of access to open data. It does so by providing certificates to each open dataset made available by the platform's users. Using a self-administered questionnaire to collect the data, the certificates document characteristics of the datasets according to dimensions such as technical, legal, practical and social.

The Open Data Certificates are visualised in report form, displaying the complete data collected via the questionnaires on each open dataset. Each dataset is also given a total categorical score icon: Raw, Pilot, Standard or Expert.

Key points relevant to OpenDataMonitor

- Uses mostly closed questioning to collect a number of **short, factual** and **self-certified** answers (rather than quantitative data) from the platform's users in a **non-automated fashion**
- The questionnaire may expand and contract, as **the quantity of metrics used is dependent on the characteristics** of the dataset in question and not all metrics will be relevant
- Focus is **not limited to a particular geographical location** and allows for the certification of **all forms of open data**
- Each certificate is **made available in its raw data form** and **open source code is used** to power the platform

2.1.7 ENGAGE

ENGAGE, developed under the European Commission FP7 Programme, seeks to act as a social platform for open data. The platform allows users to submit and maintain open data sets using its data management system and now acts as a portal for over 52,000 datasets. Users may also utilise the public sector datasets stored on the platform. The platform uses a number of features to stimulate contact and collaboration between users, which fosters an open data community. In addition to the provision of the qualitative features of the datasets and users, ENGAGE also provides basic dataset statistics.

The ENGAGE platform allows for users to browse and search datasets based upon attributes such as date, popularity, geography, licensing and format. It also encourages user-generated visual content in the form of graphs and tables, which are then attached to the dataset on the platform for other users to view and use.

Key points relevant to OpenDataMonitor

- Social portal platform uses **6 qualitative** and **3 quantitative metrics** to measure the features of the uploaded datasets
- Methodology is **non-automated** and **self-certified** as users are able to upload and manage their own data sets
- Adopts an **EU focus only** although does allow for the upload of both **public and private sector open data**
- **Raw data is made available** but the **code is unspecified and unavailable**

2.2 Gold-standard characteristics of open data

As the previous examples indicate, there are numerous existing attempts to monitor aspects of the open data ecosystem. In developing a comprehensive set of metrics for the OpenDataMonitor project it is worth considering the characteristics of open data that may be considered the gold-standard for measurement, barring any practical considerations.

In 2007 a group of open government advocates drafted a set of eight principles of open government data (OGD).¹ The list, taken from their website, is reproduced below including a short description of each principle.

Table 1. Eight principles of open government data

Principle	Description
1. Complete	All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. Primary	Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. Timely	Data is made available as quickly as necessary to preserve the value of the data.

¹ <http://opengovdata.org>

4. Accessible	Data is available to the widest range of users for the widest range of purposes.
5. Machine processable	Data is reasonably structured to allow automated processing.
6. Non-discriminatory	Data is available to anyone, with no requirement of registration.
7. Non-proprietary	Data is available in a format over which no entity has exclusive control.
8. License-free	Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

For practical reasons, adherence with all of these principles can not be assessed in an automated fashion. Table 2 briefly discusses these principles in more detail and explores their potential as a guide to benchmarking the open data ecosystem.

Table 2. Feasibility of measuring the eight principles

Principle	Measure
1. Complete	<p>Completeness may be measured automatically, however, any metric has to be reviewed over time. The set of open data evolves as we gain further understanding of its impact and usefulness. It may be possible to compare completeness against a pre-defined universe of open data. However, the assessment of what could be open is highly subjective and potentially a manual process. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Frequency of catalogued datasets • Frequency of catalogues by sector of publishing organisation
2. Primary	<p>Primary relates to the source of the data. What level of aggregation is appropriate, how to define the original source, or indeed how to assess the “rawness” of data are difficult questions beyond automatic metrics. Thus, the ODM project only covers a few aspects of this principle. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Total number of catalogues • Proportion of dataset distributions in each catalogue that are not listed in any other catalogues
3. Timely	<p>We can measure up-to-date catalogues and timely data automatically provided the metadata is standardised. There are implicit arbitrary decisions such as the appropriate update frequency. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Median days since latest dataset update • Frequency of datasets with stated update frequency
4. Accessible	Accessibility can be automated for many technical aspects. For example,

	<p>the distribution of data formats or the number of languages in a catalogue are usually easy to measure. Other, perhaps social aspects, are more difficult to quantify and also depend on the scope. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Frequency of dataset distributions with previews • Frequency of different languages
5. Machine processable	<p>It is fairly straightforward to assess all individual datasets (and their associated distributions) on the extent to which they are machine-readable. However, many details may require manual input and/or only emerge as problematic in an actual application. For example, the metadata may be machine-readable on a basic level but not include a meaningful schema. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Frequency of dataset distributions that are machine-readable • Frequency of error and warnings generated by CSVlint (for CSV files)
6. Non-discriminatory	<p>If each data catalogue includes an appropriate piece of information regarding its access terms, and these terms are standardised e.g. on a national level, it may be possible to measure the extent to which open data is available without discrimination. On a pragmatic level it may be too difficult or trivial if no catalogue has any restrictions. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • n/a
7. Non-proprietary	<p>Measuring the range of data formats is usually feasible in an automated fashion. Rankings of the “openness” of different formats have been suggested, for example, with Tim Berners-Lee’s <i>5 star open data</i>. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Frequency of catalogues using specific software platforms • Frequency of dataset distributions by file format
8. License-free	<p>If each dataset includes an appropriate piece of information regarding its licence, and the number of licences is limited, it may be possible to measure the extent data is available with an open licence. Potential metrics to be included in the ODM project:</p> <ul style="list-style-type: none"> • Frequency of dataset distributions with an explicitly set license • Frequency of datasets distributions with an open license

3 Metrics and key figures for OpenDataMonitor

Central to the concept of OpenDataMonitor is a set of metrics – “key figures” – whose values give rich insights into the state of the art and evolution of open data deployment. In the following sections we present and parameterise that set of metrics developed for the OpenDataMonitor platform, such that they may be implemented in the course of the project.

The metrics presented here reflect varying levels of granularity, from those that capture aggregate features of all catalogues being monitored, to metrics related to specific geographies, to those capturing features of a specific catalogue or a specific dataset.

When considering which characteristics of open data deployment to monitor, it is critical to recognise that while many features exist that are of interest, not all will be measurable or have data available to enable monitoring. These challenges are demonstrated with a set of example characteristics in Figure 1. Mapping classes of open data metric to preliminary dimensions of availability and measurability below.

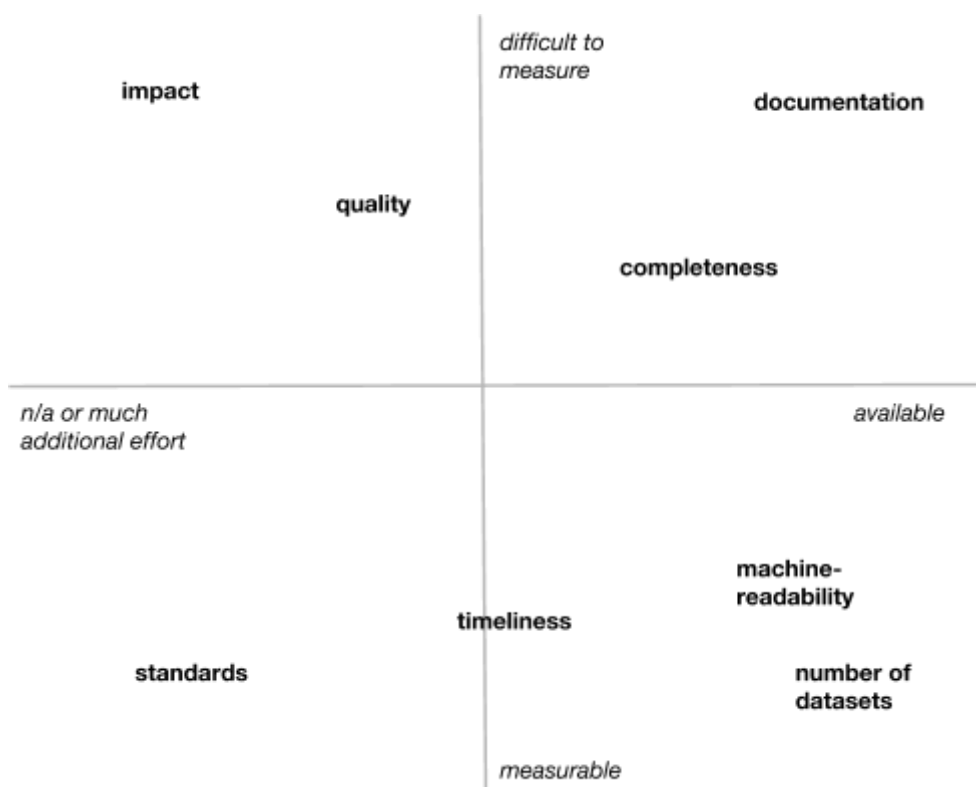


Figure 1. Mapping classes of open data metric to preliminary dimensions of availability and measurability

For example, the impact created by a particular dataset is hard to measure in objective, universal terms, and such attempts are rarely supported by available data about how a particular dataset is being used. In contrast, the machine-readability of a dataset (or its distributions) is both easier to define, based on accepted notions of the concept, and easier to assess based, for example, on metadata elements widely available in open data catalogues (e.g. the file format in which the data is encoded) or by automated inspection of the data.

Therefore, the set of metrics presented here is derived from those that are both measurable, in an objective sense, and computable based on available data, as represented by the lower right portion of Figure 1.

While it makes immediate sense to have a set of indicators that measure the development of the aggregate open data ecosystem, lower levels of aggregation may also prove useful. We grouped the metrics according the following levels: overall, per-geography, per-catalogue, and per-dataset. Some indicators start at the lower units of analysis, e.g. the dataset, and are then rolled up into higher levels.

Development of these metrics is informed both by the eight principles of open government data (see previous section) and best practices encoded in the Open Data Certificates platform and supporting questionnaires². The Open Data Certificates emphasise the importance of quality metadata in open data publishing (in addition to the quality of the data itself). It should also be noted that the metrics presented here are chosen in a trade-off between what is feasible and meaningful to measure, and an ideal set of automatic indicators.

Readers should note that the terms *catalogue*, *dataset*, and *distribution* are used in the text below as per their definitions in the DCAT vocabulary³ and discussed in OpenDataMonitor deliverable 2.1.

3.1 Measures over the aggregate

Table 3. Counts / averages / longitudinal

Tag	Label / Notes	Unit	Data type	Number of Values	Range Per Value	Sum of Values
catfreq	Total number of catalogues	n	int	1	> 0	-
catsect freq	Frequency of catalogues by sector of publishing organisation Values: <ul style="list-style-type: none"> ● government/public sector ● other non-commercial/third sector ● commercial ● mixed 	n	int	many (4)	>= 0	-
catsect prop	Proportion of catalogues by sector of publishing organisation Values: <ul style="list-style-type: none"> ● government/public sector ● other non-commercial/third sector ● commercial ● mixed 	%	float	many (4)	0-100	100

² <https://certificates.theodi.org>

³ <http://www.w3.org/TR/vocab-dcat/>

catsoft freq	Frequency of catalogues using specific software platforms Values: <ul style="list-style-type: none"> CKAN, Socrata, etc; see D2.1 Section 4.2 for indicative list 	n	int	TBC	≥ 0	-
catsoft prop	Proportion of catalogues using specific software platforms Values: <ul style="list-style-type: none"> CKAN, Socrata, etc; see D2.1 Section 4.2 for indicative list 	%	float	TBC	0-100	100
catme dageye ars	Median age of catalogues Note: use the date of publication of the first dataset as a proxy for the data of the catalogue launch	years	fixed (2)	1	> 0	-
catme anagey ears	Mean age of catalogues Note: use the date of publication of the first dataset as a proxy for the data of the catalogue launch	years	fixed (2)	1	> 0	-
catne wmont hfreq	New catalogues per month Note: use the date of publication of the first dataset as a proxy for the data of the catalogue launch; either use an arbitrary start date before any catalogues were launched, or the “launch” date of the first catalogue as the start date.	n	int	> 0	≥ 0	-

Additional notes for the metrics outlined in Table 3:

- The **total number of catalogues** is a straightforward count. We urge any interpretation to go beyond “more is better” and any visualisation and/or report ought to reflect how quality of catalogues may be more important than quantity.
- **Frequency (proportion) of catalogues by sector of publishing organisation** describes the count of catalogues by its predominant sector. Open data is not only about open government data; hence we are bringing awareness to non-governmental sources. For example, universities are a substantial non-governmental, non-commercial source of open data (both research and operational data). The category “mixed” captures other cases or those where there is no clear prevailing category. It may involve some manual intervention as a one-off starting point.
- **Frequency (proportion) of catalogues using specific software platforms.** This metric should be computed to include the six software platforms described in Section 4.2 of D2.1. Inclusion

of CKAN and Socrata should be considered a minimum requirement, but limiting the scope to less than the six platforms discussed in D2.1 would present a credibility or bias issue. Some manual intervention may be required to source lists of instances of less widely used platforms.

- **New catalogues per month** uses the date of publication of the first dataset as a proxy for the data of the catalogue launch. We may either use an arbitrary start date before any catalogues were launched (e.g. 2005), or the “launch” date of the first catalogue as the start date. This may feed into a more sophisticated metric, such as the “maturity rating” in *benchmarking open government* (Veljković et al. 2014)⁴. The rationale for a maturity model lies in the putative upper bound of the number of catalogues.

Table 4. Rankings

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catfreq ranktop	Highest frequency of catalogues per country	rank	int	10	1-10	-
catfreq rankbottom	Lowest frequency of catalogues per country	rank	int	10	1-10	-
catcapit ranktop	Highest frequency of catalogues per capita per country Data Source: http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD	rank	int	10	1-10	-
catcapit rankbottom	Lowest frequency of catalogues per capita per country Data Source: http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD	rank	int	10	1-10	-

Additional notes for the metrics outlined in Table 4:

- The top and bottom ranks for various metrics are useful for benchmarking. It is not *a priori* clear whether the top countries/regions/etc represent a meaningful distinction, however, we will gain additional insights into the open data ecosystem.

⁴ Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*. doi:10.1016/j.giq.2013.10.011, pp. 8

- Note that the two metrics above are a starting point and may be extended with many more rankings, computed in the same manner. Further rankings may include:
 - top/bottom 10 catalogues by proportion of dataset distributions that are properly licensed
 - top/bottom 10 countries by proportion of dataset distributions that are properly licensed
 - top/bottom 10 catalogues by total data size
 - top/bottom 10 countries by total data size
 - top/bottom 10 catalogues by proportion of data file links that are broken
 - top/bottom 10 countries by proportion of data file links that are broken

3.1.1 Summary statistics based on lower levels of aggregation

Many metrics collected on a lower level, e.g. size for each dataset distribution, can be aggregated and presented at a higher level such as total data size per catalogue or even country. We are not listing all viable metrics individually because of space and therefore present a template of which descriptive statistics are most useful. The complete distribution, e.g. in the form of a histogram, may also be a potential visualisation.

If a metric is aggregated via descriptive statistics the following default template applies. An example of applying this template is given in the 'Data volumes' section of 'Per-catalogue measures', specifically the metrics named `catdatasize[*]`.

Table 5. Templates for summary statistics for aggregated lower-level metrics

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
cat__total	Total (sum)	n	float	1	> 0	-
cat__med	Median	n	float	1	> 0	-
cat__mean	Mean	n	float	1	> 0	-
cat__min	Minimum	n	float	1	> 0	-
cat__max	Maximum	n	float	1	> 0	-
cat__stddev	Standard deviation	n	float	1	>= 0	-

There is also a case for aggregate measures that reflect the state of the art across all catalogues in the sample, based on lower-level metrics. For example, it may be of interest to compute the mean/median of *catlicensedprop* (proportion of distributions with an explicitly set license) across all catalogues. The implementation cost of doing so would be trivial once the data is gathered and initial metrics computed. Therefore it is proposed that such aggregate metrics are implemented in a later iteration based on identified user needs/feedback.

3.2 Per-geography measures

Table 6. Catalogue statistics across geographical regions

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catgeofreq	Catalogues per geographic region Values: <ul style="list-style-type: none"> Country State Region City/Locality Please see notes!	n	int	TBC	>= 0	-
catcapitafreq	Catalogues per capita per country Data source: http://data.worldbank.org/indicator/SP.POP.TOTL	n	int	28 EU member states	>= 0	-
catgdpcorr	Catalogues & per-capita GDP correlation (Pearson and Spearman's rank) Data source: http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD	n	float	28 EU member states (times 2)	(-1)-1	-
cathdicorr	Catalogues & HDI correlation (Pearson and Spearman's rank) Data source: https://data.undp.org/dataset/Table-1-Human-Development-Index-and-its-components/wxub-qc5k	n	float	28 EU member states (times 2)	(-1)-1	-
catcountrysectfreq	Frequency of catalogues by sector of publishing organisation Values: <ul style="list-style-type: none"> government/public sector other non-commercial/third sector commercial mixed 	n	int	many (4)	>= 0	-
catcountrysectprop	Proportion of catalogues by sector of publishing organisation Values: <ul style="list-style-type: none"> government/public sector other non-commercial/third sector commercial mixed 	%	float	many (4)	0-100	100

Additional notes for the metrics outlined in Table 6:

- **Catalogues per geographic region** are a set of metrics that count the frequency of catalogues per country and smaller geographic region. This means that
 - each EU member state has a total count;
 - each smaller geographic unit, e.g. modelled after the EU Nomenclature of Territorial Units for Statistics (NUTS) 1, 2, and 3, has a total count; each EU member state has a “family tree”, where we model the high-level admin geography using the values above, down to city level (as the smallest unit) and then map catalogues to each level in the hierarchy of catalogues within that geography.
- **Catalogues per capita per country** is a relative measure that accounts for population. Again, this is intended as an exploratory metrics and its interpretation must not rely on a ‘highest = best’ rating.
- **Catalogues & per-capita GDP (HDI) correlation (Pearson and Spearman’s rank)** are also intended as exploratory metrics.
- **Frequency (proportion) of catalogues by sector of publishing organisation** describes the count of catalogues by its predominant sector. Open data is not only about open government data, hence we are bringing awareness to non-governmental sources. For example, universities are a substantial non-governmental, non-commercial source of open data (both research and operational data). The category “mixed” captures other cases or those where there is no clear prevailing category. It may involve some manual intervention as a one-off starting point.

Table 7. Time-profile by country

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catcount rynewm onthfreq	New catalogues per country per month Note: use date of first dataset as proxy for catalogue launch	n	int	28 EU member states (times months)	>= 0	-

It should be noted that as additional longitudinal data is collected by the OpenDataMonitor platform, it may be feasible and desirable to collect additional time-profile metrics, for example the change over time in terms of data formats used for data publication.

3.3 Per-catalogue measures

Table 8. Data volumes

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catdatasetsfreq	Frequency of catalogued datasets	n	int	1	>= 1	-
catdistributionsfreq	Frequency of catalogued distributions	n	int	1	>= 1	-
catpublishersfreq	Frequency of unique organisations publishing data See additional notes.	n	int	1	>= 1	-
catdatasettotal	Total distribution size in a catalogue	KB	float	1	>= 0	-
catdatasetmed	Median distribution size	KB	float	1	>= 0	-
catdatasetmean	Mean distribution size	KB	float	1	>= 0	-
catdatasetmax	Maximum distribution size	KB	float	1	>= 0	-
catdatasetstddev	Standard deviation of distribution sizes	KB	float	1	>= 0	-

Additional notes for the metrics outlined in table 8:

- **Frequency of catalogued datasets** is a count of the number of distinct datasets. For example, a dataset could be a series of expense claim data for the same department published monthly. This metric gives a broad indication of the amount of open data available.
- **Frequency of catalogued distributions.** Distributions are the individual data files containing the data. A dataset may have zero or more associated distributions. For example, each month's expense claims mentioned above may be detailed in a separate CSV file; each of these would comprise a unique distribution, all associated with the same dataset as a logical grouping. This metric gives an indication of the extent and granularity of open data available.
- **Frequency of unique organisations publishing data** counts the number of organisations that are separate entities. The metric is based on values of the CKAN schema, as mapped out in D2.2, namely *dct:publisher*, and similar metadata. This metric may require some data

processing for deduplicating names if spellings differ. Overall, the aim is to measure the breadth of data publishers.

- Descriptive statistics for **distribution size** are calculated from the individual values of each distribution. Thus, it is an application of the previous section: *summary statistics based on lower levels of aggregation*.

Table 9. Data duplication/uniqueness

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catdupl prop	Proportion of distributions in each catalogue that are <i>listed</i> in other catalogues	%	float	1	0-100	-
catuniq prop	Proportion of distributions in each catalogue that are <i>not listed</i> in any other catalogues	%	float	1	0-100	-

Additional notes for the metrics outlined in table 9:

- The **proportion of distributions in each catalogue that are listed in other catalogues** (and its inverse) are a measure to understand how many distributions are syndicated in other, e.g. national, catalogues. This metric will provide
 - context for the frequency counts of distributions in catalogues;
 - a way of adjusting/extending the frequency counts to take into account unique distributions.
- Pseudo-code to implement these metrics may appear as follows:

```

download and checksum all the distributions pointed to
build index "ix_checksums": checksum -> catalogues
build a second "ix_catalogues": catalogue -> checksums

for each catalogue in ix_catalogues
  var duplicated = 0
  checksums = ix_catalogues[catalogue]
  for each checksum in checksums
    if ix_checksums[checksum].size > 1
      duplicated++
    end;
  end;
end;
```

Table 10. Housekeeping

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catbrokenlinksprop	Proportion of data file links that are broken Note: broken links defined by HTTP response codes in the 4xx and 5xx range, with exceptions, e.g. 400	%	float	1	0-100	-
catstatcodeprop	Proportion of different HTTP status codes for data file URIs Values: see http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html	%	float	many	0-100	100

Table 11. Formats and machine-readability

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catfileformatfreq	Frequency of distributions by file format Values: [Compressed/Archive (zip, gz, tar, rar), HTML, PDF, Word, Excel, CSV, TSV, JSON, XML, RDF (all serialisations), other or unknown Notes: based on catalogue metadata	n	int	many (11)	>= 0	-
catfileformatprop	Proportion of distributions by file format Notes: as above for frequency.	%	float	many (11)	0-100	100
catmachinereadformatfreq	Frequency of distributions in a machine-readable file format Notes: <ul style="list-style-type: none"> where machine readable means one of: CSV, TSV, JSON, XML, RDF (all serialisations) based on catalogue metadata 	n	int	1	>= 0	-
catmachinereadformatprop	Proportion of distributions in a machine-readable file format Notes: as above for frequency.	%	float	1	0-100	-
catmimet	Frequency of distributions by MIME type of	n	int	many	>= 0	-

yepfreq	data file Indicative values (non-exhaustive): <ul style="list-style-type: none"> • application/x-zip-compressed • application/pdf • text/html • text/csv • application/json • application/xml • application/rdf+xml Notes: list of registered media types: http://www.iana.org/assignments/media-types/media-types.xhtml					
catmimet ypeprop	Proportion of distributions by MIME type of data file Notes: as above for frequency.	%	float	many	0-100	100
catmachi nereadfr eq	Frequency of distributions that are machine-readable Counted as machine-readable if: <ul style="list-style-type: none"> • CSV/TSV: contains a header row that starts in row 1 and data in row 2. OR no header and data in row 1. • JSON: error-free native parser request. • XML: error-free native parser request. • RDF: error-free native parser request. 	n	int	many (4)	>= 0	-
catmachi nereadpr op	Proportion of distributions that are machine-readable Notes: as above for frequency.	%	float	many (4)	0-100	100

Additional notes for the metrics outlined in Table 11:

- **Proportion**, for each metric, is the relative metric (as a percentage) based on the frequency.
- **Frequency of distributions by file format** is a count of the various data formats available in a catalogue. Data formats that are not listed fall into another category. If the frequency of 'other' surpasses a substantial threshold, it may be worth introducing further categories.
- **Frequency of distributions in a machine-readable file format** counts the number of formats deemed machine-readable. This metric is only a count based on the metadata available and does not evaluate the actual files.
- **Frequency of distributions by MIME type of data file** is another way of measuring the variety in file formats and the quality of metadata.
- **Frequency of distributions that are machine-readable** evaluates each file with an algorithm to check whether the file is available and machine-readable. Some considerations are based

on a ODI study on CSV in data.gov.uk from February 2014⁵. Standardised formats like XML or RDF can be tested with a parser. CSV is more difficult because it has few specifications. In practical terms a machine-readable dataset ought to have a header row and data in the second row. Sometimes the header row may be omitted, but all other variations (e.g. source in the first three rows) require guessing or human input.⁶

Table 12. Licenses

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catlicens edfreq	Frequency of distributions with an explicitly set license Notes: based on the catalogue metadata	n	int	1	>= 0	-
catlicens edprop	Proportion of distributions with an explicitly set license Notes: based on the catalogue metadata	%	float	1	0-100	-
catopen licfreq	Frequency of distributions with an open license Notes: <ul style="list-style-type: none"> based on the catalogue metadata list of open licenses: http://opendefinition.org/licenses/ 	n	int	1	>= 0	-
catopen licprop	Proportion of distributions with an open license (excluding and including distributions with missing licenses) Notes: <ul style="list-style-type: none"> based on the catalogue metadata list of open licenses: http://opendefinition.org/licenses/ 	%	float	2	0-100	-
catdsbyl icensefr eq	Frequency of distributions by license type Notes: <ul style="list-style-type: none"> based on the catalogue metadata include all licenses found in the metadata, irrespective of openness 	n	int	many	>= 0	-
catdsbyl icensepr op	Proportion of distributions by license type (excluding and including distributions with missing licenses)	%	float	many (times 2)	0-100	100

⁵ We analysed more than 20,000 links to CSV files on data.gov.uk – only around one third turned out to be machine-readable. <http://theodi.org/blog/the-status-of-csvs-on-datagovuk>

⁶ Another available CSV tool from the ODI is <http://csvlint.io>.

	Notes: <ul style="list-style-type: none"> based on the catalogue metadata include all licenses found in the metadata, irrespective of openness 					
--	------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--	--	--	--	--

Additional notes for the metrics outlined in Table 12:

- Proportion**, for each metric, is the relative metric (as a percentage) based on the frequency.

Table 13. Release frequencies and timeliness

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catmedsi nceupdat edays	Median days since latest dataset update	days	int	1	>= 0	-
catmedsi ncenewd ays	Median days since latest new dataset	days	int	1	>= 0	-
catlastup datebyye arfreq	Frequency of dataset last update by year	n	int	many	>= 0	-
catupdat efreqfreq	Frequency of datasets with stated update frequency	n	int	1	>= 0	-
catupdat efreqpro p	Proportion of datasets with stated update frequency	%	float	1	0-100	-
cattau	<i>Tau</i> of the catalogue Notes: defined in http://project.opendatamonitor.eu/images/downloads/OpenDataMonitor_Publication_The-Tau-of-Data.pdf	n	float	1	0-1	-

Additional notes for the metrics outlined in the table above:

- Note that additional time-based metrics could be added here, e.g. max/min time length since last update, proportion of datasets updated per year, but these would have limited value due to the differing periodicity of datasets. For example, a census may be updated once per decade, while air quality data may be updated on a per-second or per-minute basis. Therefore it is desirable to defer to the more sophisticated metric *tau*, featured above.
- The ***tau* of the catalogue** is the percentage of datasets up-to-date in a data catalogue. To calculate timeliness, and ultimately the tau, two metrics are required: the last substantial

update and a standardised update frequency for each datasets. We recommend, similar to the paper using the default values of a delta (δ) of one day and lambda (λ) of 10 percent.

Table 14. Prominence, engagement and usability

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
catsitepagerank	PageRank of the catalogue site Notes: <ul style="list-style-type: none"> this should be computed for the “pay-level domain”, e.g. data.gov.uk not gov.uk potential library for retrieving pagerank scores is here: https://github.com/eyecatchup/SEOstats 	n	int	1	0-10	-
catuniquepublishersfreq	Frequency of unique publishers contributing to the catalogue Notes: <ul style="list-style-type: none"> based on CKAN metadata “maintainer” or “author” 	n	int	1	≥ 0	-
catuniquepublishersprop	Frequency of unique publishers relative to catalogue size Method: value = unique publishers / datasets in catalogue	n	float	1	0-1	-
catapisdumpsfreq	Frequency of datasets available via APIs and/or data dumps Notes: <ul style="list-style-type: none"> a data dump is defined as a file (or files) containing the entire data set, with no specialised query mechanism required some datasets may provide both an API and data dumps 	n	int	1	≥ 0	-
catapisdumpspop	Proportion of datasets available via APIs and/or data dumps Notes: <ul style="list-style-type: none"> a data dump is defined as a file (or files) containing the entire data set, with no specialised query mechanism required some datasets may provide both an API and data dumps 	%	float	1	0-100	-

catapisdu mpsratio	Ratio of datasets with APIs to those with data dumps	n	float	1	>= 0	-
catdspre viewsfreq	Frequency of distributions with previews	n	int	1	>= 0	-
catdspre viewspro p	Proportion of distributions with previews	%	float	1	0-100	-
catlangs	Frequency of different languages	n	int	1	>= 1	-

Additional notes for the metrics outlined in Table 14:

- **Proportion**, for each metric, is the relative metric (as a percentage) based on the frequency.

3.4 Per-dataset measures

Table 15. Data and metadata volume, quality and usability

Tag	Label / Notes	Unit	Data type	Number of Values	Range per Value	Sum of Values
dssize	Dataset size Notes: <ul style="list-style-type: none"> • useful as a descriptive statistic • computed as the sum of the bytesize of all distributions associated with the dataset 	KB	float	1	>= 0	-
dspopul atedmdf ields	Number of fields in the metadata record that are populated	n	int	1	>= 1	-
dsvocab susedfr eq	Frequency of unique vocabularies used in metadata record Notes: <ul style="list-style-type: none"> • including the catalogue's native vocabulary 	n	int	1	>= 1	-
dsvocab termsus edfreq	Frequency of terms used from each vocabulary present in metadata record Notes: <ul style="list-style-type: none"> • including the catalogue's native vocabulary 	n	int	>= 1	> 0	-
dsvocab termsus	Proportion of terms used from each vocabulary present in metadata record	%	float	>= 1	1-100	100

edprop	Notes: <ul style="list-style-type: none">including the catalogue's native vocabulary					
dsodcertlevel	Open Data Certificate level of the dataset Values: <ul style="list-style-type: none">- Raw- Pilot- Standard- Expert Notes: See https://certificates.theodi.org/ for more information	-	enum	1	-	-
dscsvvalidationfreq	Frequency of Errors and Warnings generated by CSVlint Notes: applies only to datasets in CSV format	n	int	2	>= 0	-
dstimeliness	Timeliness of the dataset Notes: <ul style="list-style-type: none">a measure of whether or not the dataset can be considered up-to-datedefined in http://project.opendatamonitor.eu/images/downloads/OpenDataMonitor_Publication_The-Tau-of-Data.pdf	bool	bool	1	0 1	-

3.5 Summary

Having defined a wide range of metrics operating at different granularities, the following sections will explore how these populated metrics may be represented graphically, either as individual visualisations or aggregated into dashboards.

4 Visualisation techniques for monitoring open data

4.1 Introduction to data visualisations

Visualisations of data, maps and concepts date back even to prehistoric times. The interested reader can explore an extensive timeline in Michael Friendly's paper *Milestones in the history of thematic cartography, statistical graphics, and data visualization* also available online.⁷

We find the earliest use of modern techniques in 1786 by Scottish engineer William Playfair with bar charts and line graphs of economic data. For many the founder of graphical methods of statistics, he also invented the pie chart and the circle graph.

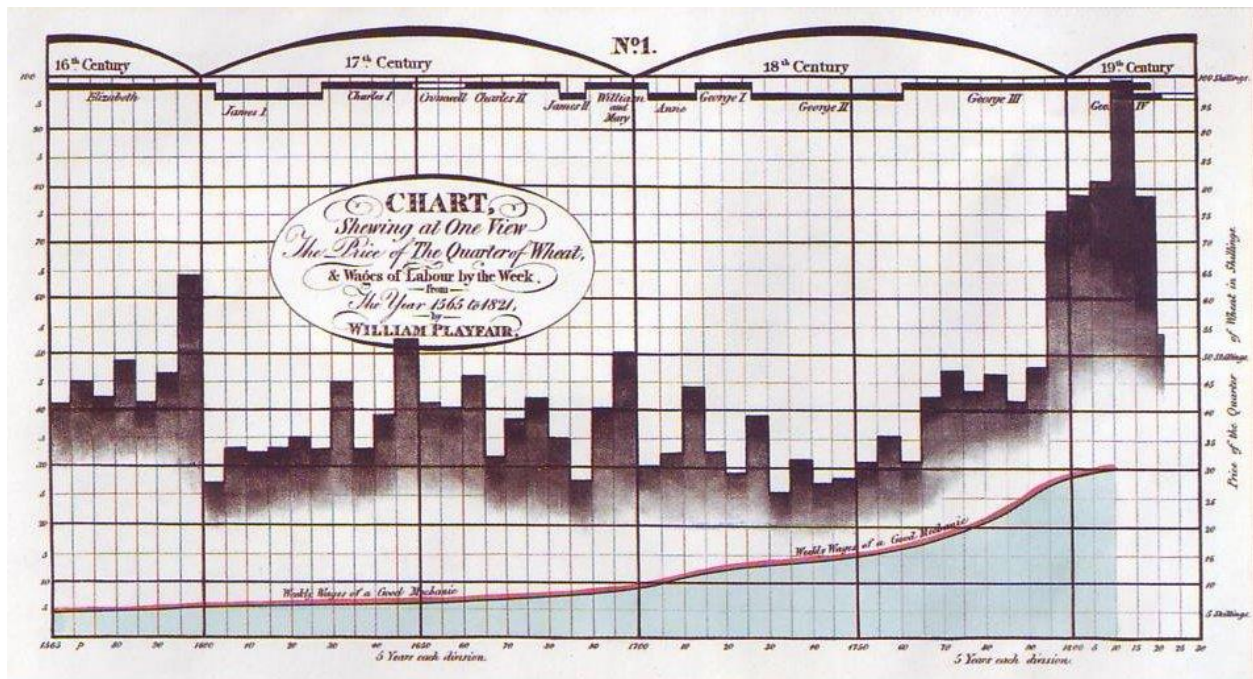


Figure 2. William Playfair's visualisation of wheat prices.

What followed soon after is what Friendly calls the "Golden Age of data graphics" from 1850-1899. Famous example such as John Snow's map of cholera deaths in London, Florence Nightingale's visual campaign to improve sanitary conditions, Quetelet and Galton's statistical contributions, Minard's map of Napoleon's March on Moscow and many more were created in this period.

The first half of the 20th century was relatively quiet, perhaps with the notable exception of Otto Neurath's *Vienna Method of Pictorial Statistics* (later known as Isotype, International System of Typographic Picture Education) developed between 1925 and 1934.

The second half of the 20th century, with the rise of computers, have lead us into a new age of visualisations. From this period stem several seminal treatments such as Jacques Bertin's *Sémiologie Graphique* (1967)⁸, Edward Tufte's *The Visual Display of Quantitative Information* (1983)⁹ and William Cleveland's *The Elements of Graphing Data* (1985).¹⁰

⁷ <http://www.datavis.ca/milestones>

⁸ Bertin, Jacques (1967) *Sémiologie Graphique: Les diagrammes, les réseaux, les cartes*. Gauthier-Villars. Paris.

The 21st century brought largely innovations for the interactive and digital space. This includes new programming languages such as the popular JavaScript library D3: data-driven documents.¹¹ Another modern example that excels in visualising many dimensions, displaying changes over time, and allowing users to explore the data interactively is Gapminder.¹²

The ODM project mainly focuses on automated graphics for online consumption. This implies that the emphasis lies on the principles of statistical graphics outlined, for example, in Tufte (1983), and less on the playful nature of infographic design. The workhorses in this context are therefore classic visualisations such as the bar and line chart. The digital medium, unlike print graphics, allows us to provide additional context and interactive elements to supplement these visualisation forms.

4.2 Guidance and principles in applying visualisation techniques

“Above all else show the data.”

Edward Tufte

How do we create exemplar visualisations? To truly answer this question we have to consider the data, the context, the audience, the consumption and so forth. However, a few general guidelines ought to apply to all visualisations. There are several seminal reference that explain how to create good information display. The following three principles are adapted from Tufte (1983) and other sources:

1. show the data,
2. emphasise the data,
3. inform and engage.

4.2.1 Show the data

Integrity

The visualisation of data should foremost be truthful: this means, firstly, that the representation of numbers is directly proportional to the quantities that are displayed. Secondly, it means that the size of the effect in the visualisation, or “story”, also represents the size of the effect in the data. Context is crucial and, similar to quotes with words, data should not be taken out of context. Often it is showing the variation in the data that brings out the clearest visualisation.

Kenneth Haemer explains a few pitfalls of presenting data in a series of *The American Statistician* (1949-1951). Some of his examples are reproduced in the following table.

⁹ Tufte, Edward R. (1983) *The Visual Display of Quantitative Information*. Graphics Press. Cheshire, CT, USA.

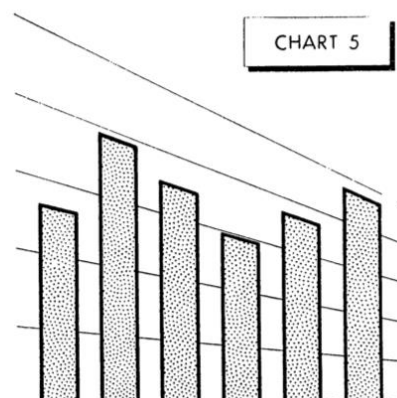
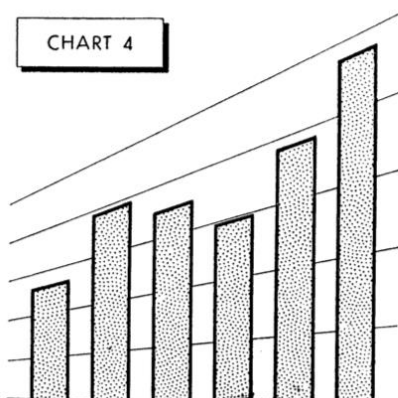
¹⁰ Cleveland, William S. (1985) *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey, USA, 1985

¹¹ <http://d3js.org>

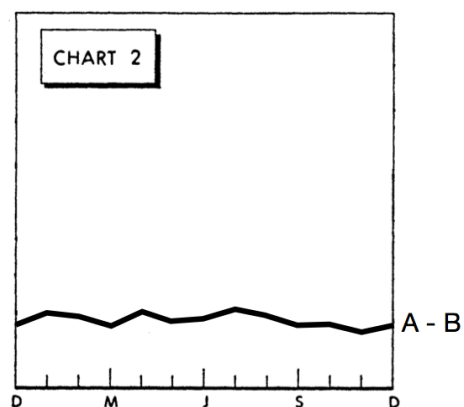
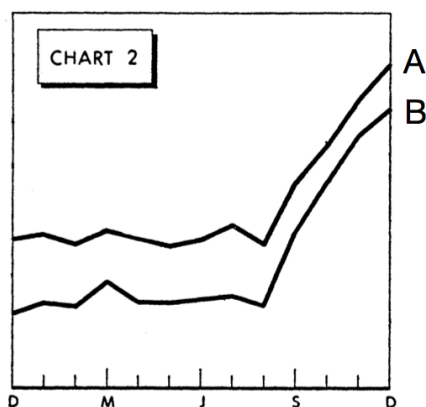
¹² Gapminder: Unveiling the beauty of statistics for a fact based world view. <http://www.gapminder.org>

Table 16. Pitfalls of data visualisations and what to avoid

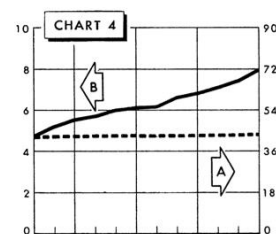
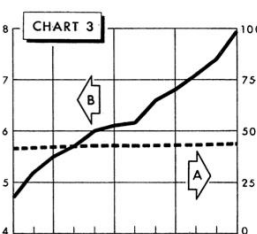
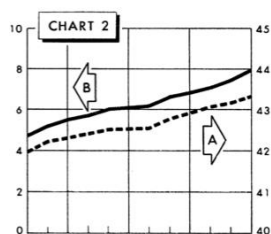
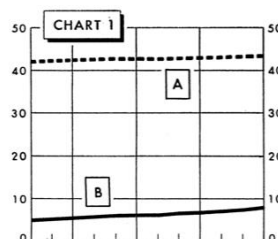
Perspective



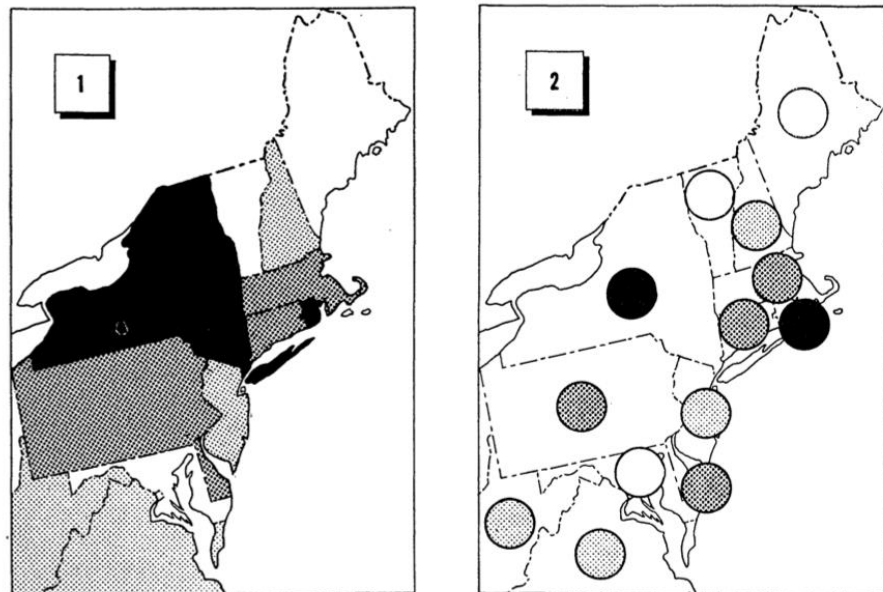
Comparing multiple lines



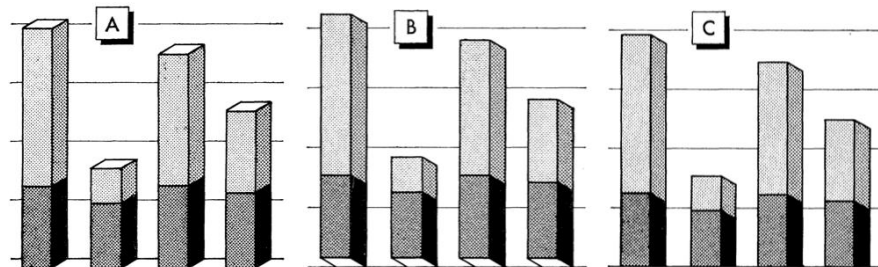
The dangers of two Y-axes



Area bias (compare also the example in section 4.4.1)



Three dimensions



Density

It is often surprising how much we can shrink a chart and still preserve the visualisation. The formula for density we find in Tufte's book is quite simple:

$$\text{Data density} = (\text{number of entries in a dataset}) / (\text{area of data graphic})$$

It can be unavoidable: as the volume of data increases, we have to either select the most relevant views, aggregate if possible or have to shrink the graphs. Two ways of achieving this, as we have seen in section 4.4.10, are small multiples and sparklines.

4.2.2 Emphasise the data

Minimise non-data ink

This guideline is a more applied version of "seek simplicity". Non-data ink may include heavy gridlines, axes, unnecessary borders and so forth. We can extend this by also erasing redundant data-ink. An example of how to go from plenty of data ink to a more parsimonious version is in the example in 4.

Avoid chartjunk

Tufte explains chartjunk as decorative elements that provide no data and may distract or confuse the viewer. Chartjunk goes against the idea of emphasising the data. This may include meaningless colours, heavy use of graphics and images, distracting embellishments etc.

4.2.3 Inform and engage

Context

The interpretation of data, even in a visualisation, crucially depends on the context. Contexts can mean the timeframe, the comparisons to other regions, people or datasets or the immediate annotations such as labels. A graphic can have two completely different meanings depending on the choices the creator makes. Unintentional, or sometimes intentional, misdirection may happen by highlighting irrelevant features of the data, comparing apples with pears, omitting critical parts and so forth.

Context varies almost by definition for each visualisation. Thus, we limit our recommendation here by suggesting to add *meaningful annotations*. You can see example in section 4.4.2 below.

Audience

Who is the audience? What is their prior knowledge? What are the needs and expectations? A great visualisation bears in mind the audience, so that it serves the aim of “inform and engage”. Somewhere in the process, perhaps after the first draft is completed, the creators should step back and imagine themselves to see the graphic for the first time. This may seem obvious, but it is easy to become your own audience.

4.3 An example of applying the principles of good visualisation design

Below is a line chart that displays the change in internet use among adult men and women in Scotland from 2001 to 2011.

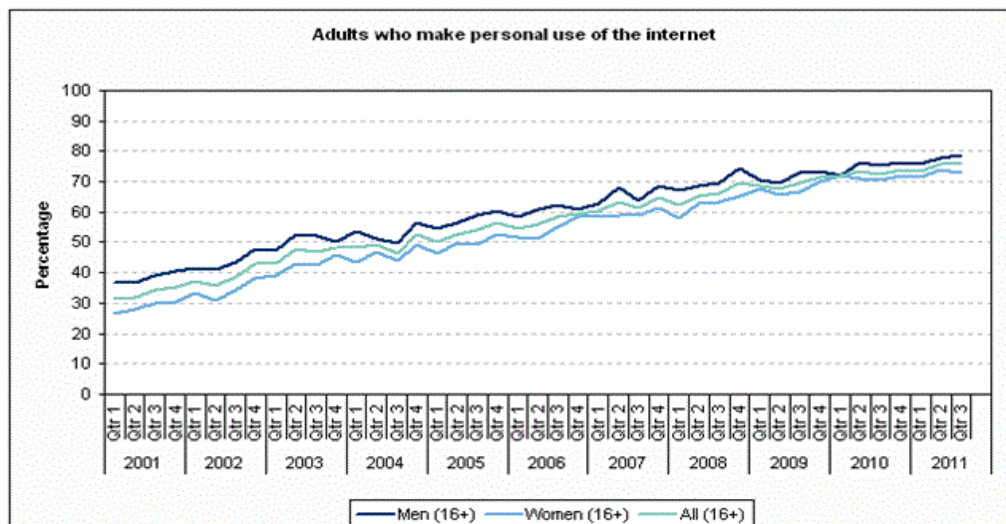


Figure 3. Internet use among adult men and women in Scotland
(Source: <http://www.scotland.gov.uk/Topics/Statistics/16002/DataTrendsInternet>)

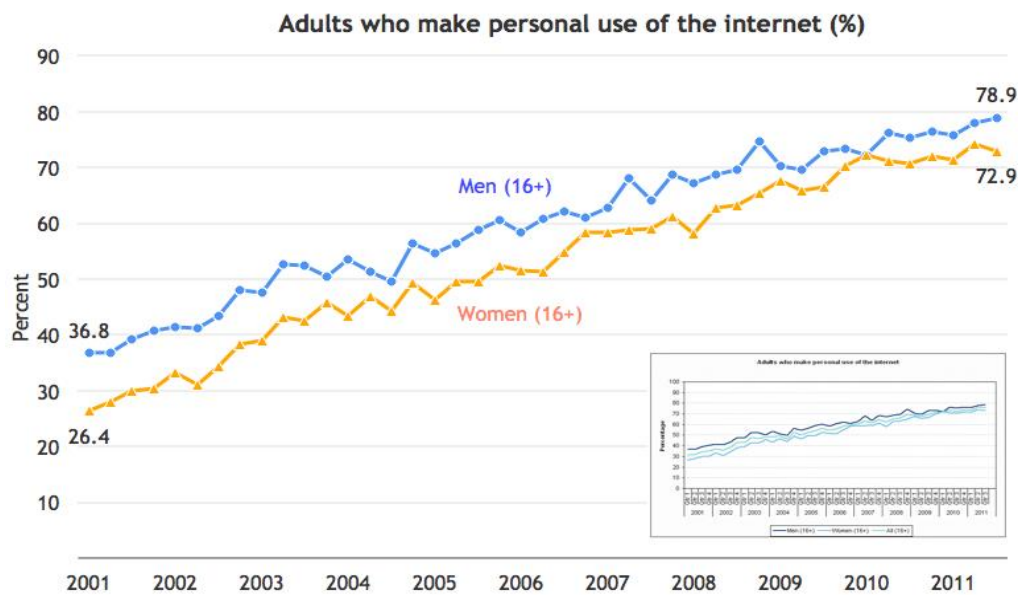


Figure 4. Internet use among adult men and women in Scotland (Redesigned from ODI)

We addressed the following points:

- Reduce clutter such as omitting the labels for individual quarters
- Minimise non-data ink by removing borders and muting gridlines
- Emphasise the data by excluding the data line for “all” as it adds no information
- Move the legend next to the lines to help the user find the appropriate reference
- Different type choice and larger font size
- Subtle use of colours that are easy to distinguish even for colour-deficient people
- Add meaningful annotations: the data labels for 2001 and 2011 tell the story of the strong uptake of internet use in one decade. (Labels in between would only add clutter; at most they can be displayed in a digital version by hovering over a data point.)
- Provide context by adding a thumbnail of the original chart

Some of these changes are inspired by what Tufte calls the “friendly data graphic”. The table is reproduced below.

Table 17. The Friendly Data Graphic¹³, Tufte (1983)

Friendly	Unfriendly
words are spelled out, mysterious and elaborate encoding avoided	abbreviations abound, requiring the viewer to sort through text to decode abbreviations
words run from left to right, the usual direction for reading occidental languages	words run vertically, particularly along the Y-axis; words run in several different directions
little messages help explain data	graphic is cryptic, requires repeated references to scattered text
elaborately encoded shadings, cross-hatching, and colors are avoided; instead, labels are placed on the graphic itself; no legend is required	obscure codings require going back and forth between legend and graphic
graphic attracts viewer, provokes curiosity	graphic is repellent, filled with chart junk
colors, if used, are chosen so that the color-deficient and color-blind (5-10% of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color-deficient people)	design insensitive to color-deficient viewers; red and green used for essential contrasts
type is clear, precise, modest; lettering may be done by hand	type is clotted, overbearing
type is upper-and-lower case, with serifs	type is all capitals, sans serif

4.4 Visualisations of the open data ecosystem

This section shows and comments a few examples of related work. We illustrate practices and techniques by example; the next section will cover a general collection of visualisations and how they can and should be applied. Visualisations are immensely popular due to their abilities to summarise and making data accessible. Our collection of visualisations covers a wide range, but is not exhaustive. The list of examples is virtually unlimited therefore we concentrate on those related to the open data ecosystem.

¹³ Tufte, Edward (1983). The Visual Display of Quantitative Information, p. 183

4.4.1 Maps

For cross-country, and within-country, comparisons a map is a natural visualisation technique. It has the advantage that most people are familiar with a map, hence, can easily interpret the data. The context is usually clear and many find a map an aesthetically pleasing visualisation.

The example in Figure 5 comes from the ePSIplatform “Europe’s one-stop shop on public sector information (PSI) re-use”. It is a visualisation of the PSI Scoreboard, a measure of the status of open data in the EU. This is a standard example of a *choropleth map*, where areas are shaded according to the data.

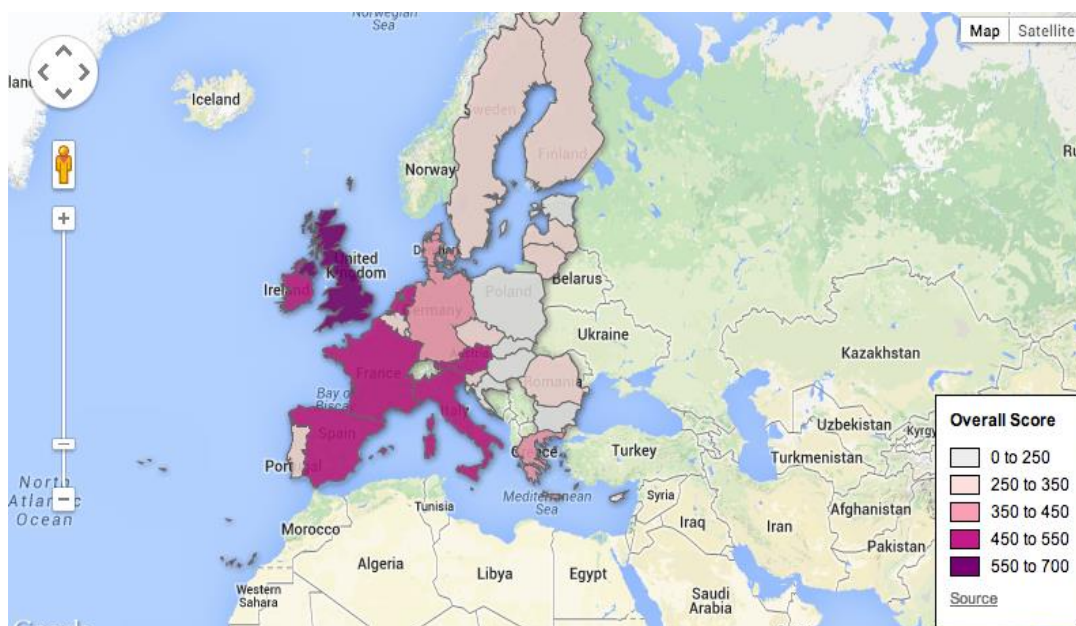


Figure 5. ePSI scoreboard
(Source: <http://www.epsplatform.eu/content/european-psi-scoreboard>)

Do	Improve
Use a limited number of colour shades	Don't neglect countries that are geographically small

The visualisation works well: notice how the colour labels are non-linear, how the shades of purple are limited and fairly easy to distinguish and how the user can explore, i.e. zoom in and out of the map.

However, it is important to realise that maps come not without shortcomings:

- A map projection is always a trade-off between the accurate size and the correct shape of a country. We recommend the Kavrayskiy VII projection.
- The area of a country does not reflect the data. For example, the Netherlands are doing well on the PSI scoreboard, but the map emphasises countries that are large such as Germany or Spain. A few solutions are presented in Table 18. Variations of the cartogram

The CTIC in Figure 6 have hinted at a solution by using symbols that represent the data on top of countries. Sometimes called a *graduated symbol map*, symbols can be scaled accordingly. This example only plots open data catalogues and is a bit too cluttered to represent leading practice.

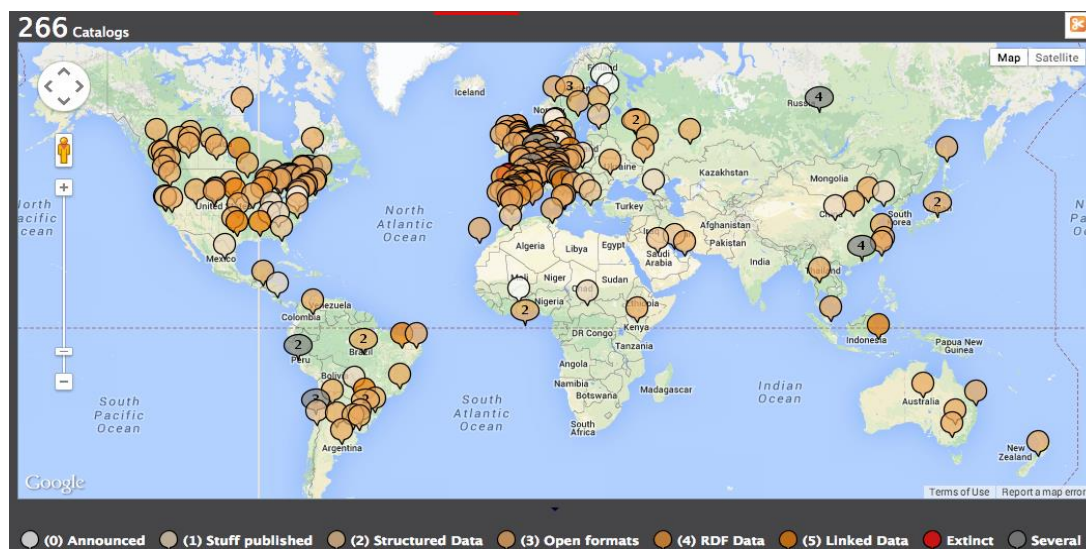


Figure 6. Graduated symbol map
(<http://datos.fundacionctic.org/sandbox/catalog/faceted>)

Do

Visualise the data by icons or symbols

Improve

Avoid cluttering the map or obfuscating data points

Another putative solution of the second mapping problem are *cartograms*¹⁴. In a project lead by the Open Data Institute, we analysed and visualised the regional geography of peer-to-peer lending in the UK. The data is available as open data. The cartogram (on the right) scales the regions of the UK according to their relative peer-to-peer activity. London is hence larger and Scotland smaller than usual.

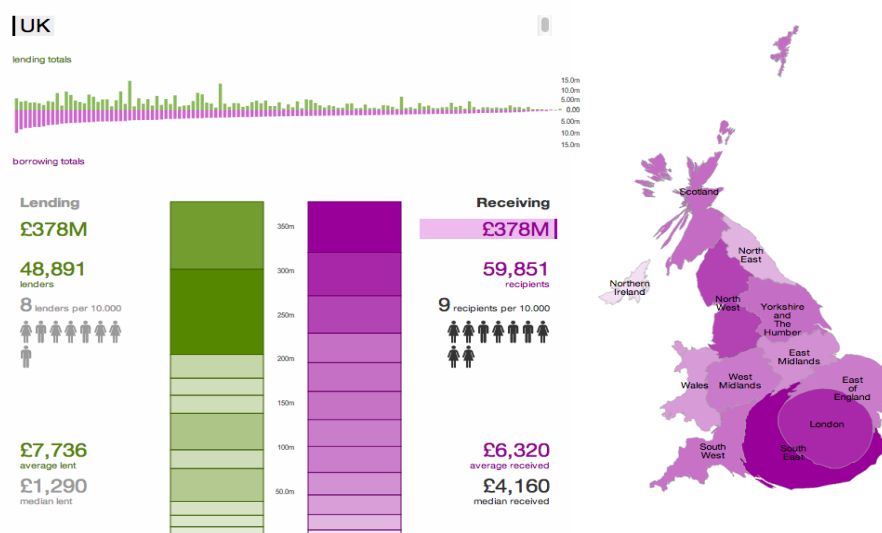


Figure 7. Cartogram: regional geography of peer-to-peer lending in the UK
(Source: <http://smtm.labs.theodi.org>)

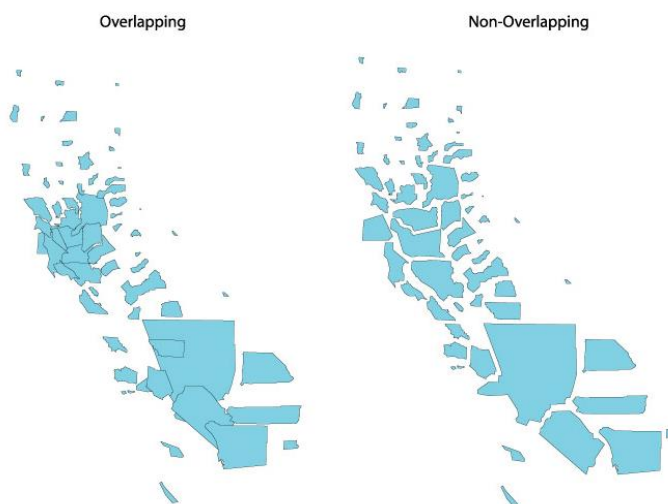
¹⁴ More than one flavour of cartogram exists, e.g. Dorling cartograms.
http://www.ncgia.ucsb.edu/projects/Cartogram_Central/cartogram_examples/dorling3.jpg

Do	Improve
The map represents the data	Remember to include contextual information

Table 18. Variations of the cartogram
(Source: http://www.ncgia.ucsb.edu/projects/Cartogram_Central/types.html)

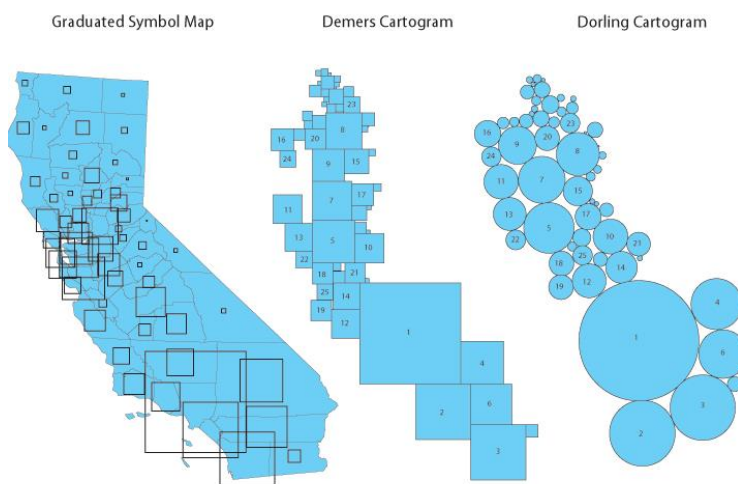
Non-contiguous cartograms

Non-Contiguous Cartograms



Dorling and Dorling-like cartograms

Dorling and Dorling-like Cartograms



4.4.2 Bar charts

Bar charts are often the optimal visualisation technique because humans can deal best with comparisons of a position along a common scale (Cleveland and McGill, 1984).¹⁵ Figure 8 shows a traditional example used for the PSI scoreboard. It is a very clean bar chart with no unnecessary borders and makes use of the digital medium by presenting more information if we hover over a specific bar (example Sweden).

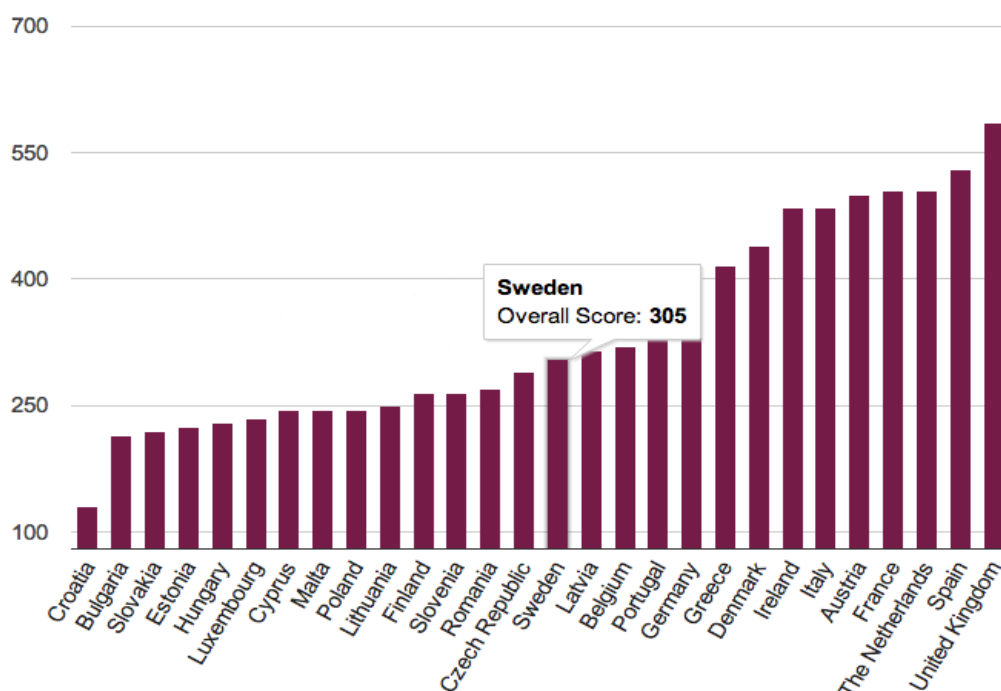


Figure 8. PSI overall score
(Source: <http://www.epsiplatform.eu/content/european-psi-scoreboard>)

Do	Improve
Clean design, additional information by mousing over the data points	Avoid annotations that are hard to read; axes usually start at zero

It could arguably be improved as a horizontal version because the labels are easier to read. The y-axis also features some non-traditional choices such as starting at 90. Below is a redesigned version that addresses these shortcomings.

¹⁵ Cleveland, W.S. and McGill, R. (1984) Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79:531–554, 1984.

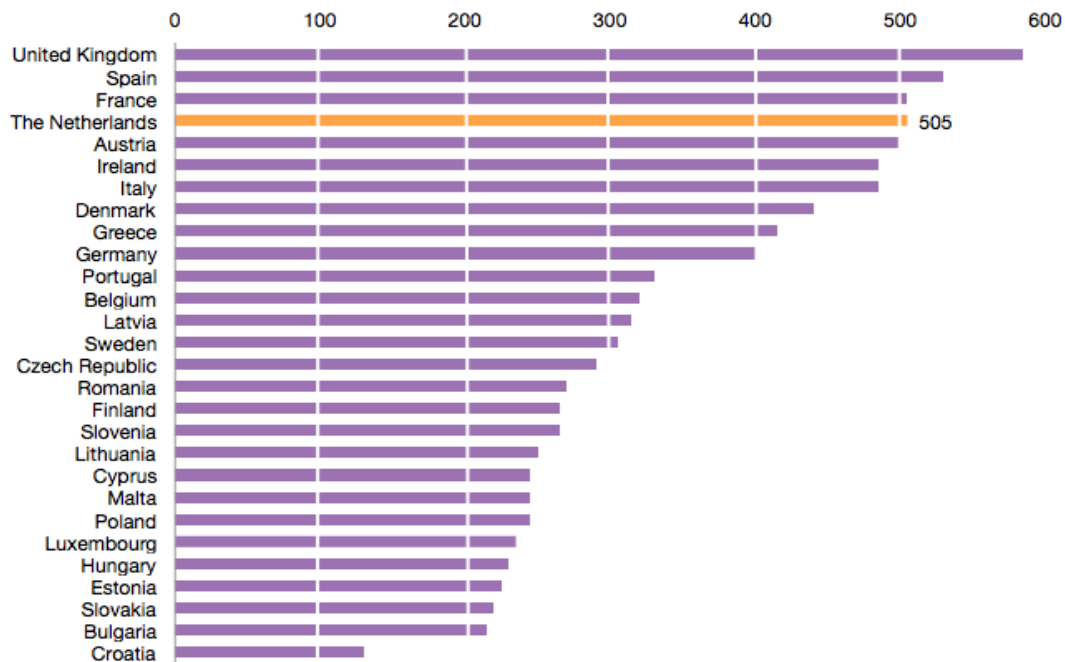


Figure 9. Redesigned bar chart with emphasis on the Netherlands (Source: ODI)

Compare this exemplar bar chart with a graphic found on the European ENGAGE project. The bar chart is animated, i.e. the bars grow upon selecting the chart. However, this does not aid the user's understanding and we find several less than optimal choices:

- The chart is missing an axis because the data labels are on top of the bar adding plenty of clutter.
- The bar labels are in the wrong place, namely not next to the bars.
- Even worse, the bar labels in the legend are in the reverse order adding an additional burden to the user (the first bar, 61, is "borough").
- Some of the data is covered by the legend.
- The colours are hard to distinguish and meaningless.

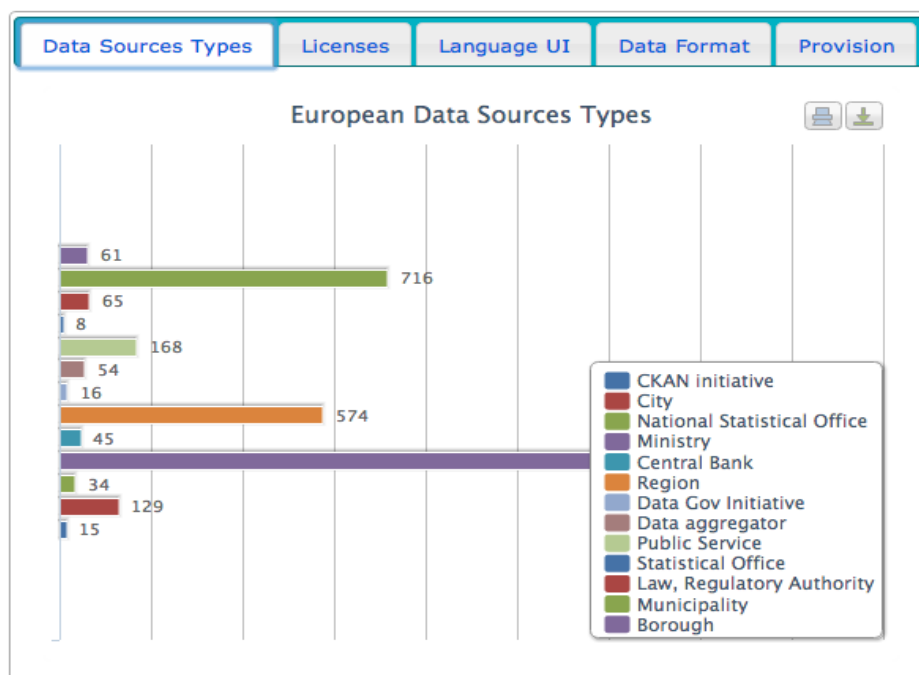


Figure 10. Example bar chart from the ENGAGE project
(Source: <http://www.engagedata.eu/opendatasites>)

Do

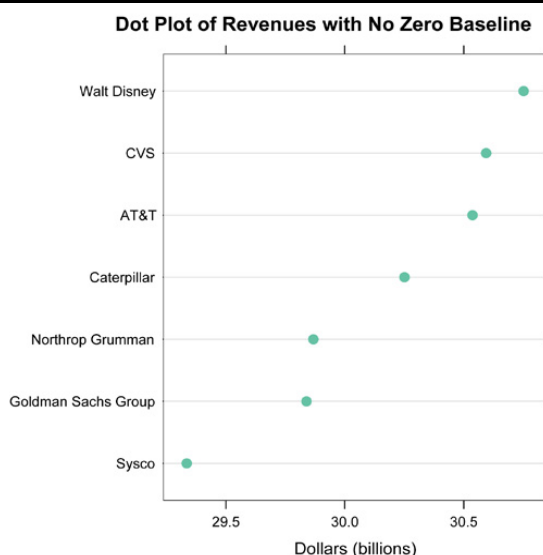
Add annotations such as bar labels that highlight the most salient information

Improve

Avoid making it difficult to see and understand the data, e.g. by having labels in the wrong place.

Table 19. Variations of the bar chart

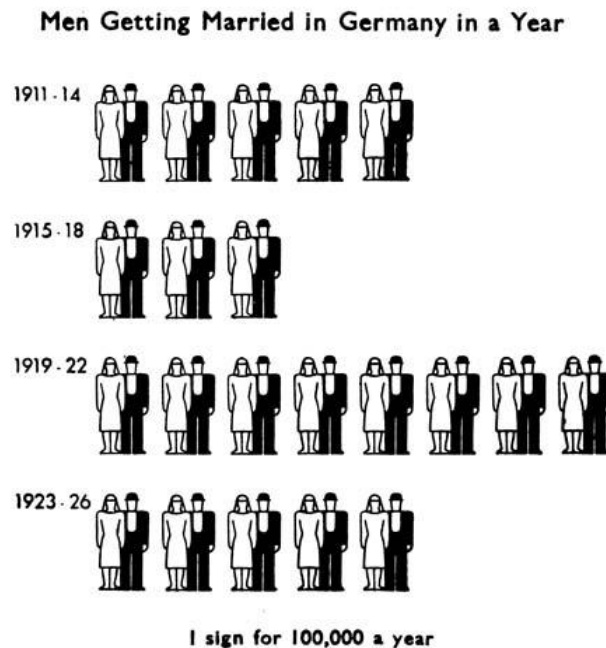
Dot plots are a useful alternative, especially when we want to truncate the axis. In this case we are still comparing the position on a common scale, but by excluding the length of the bar we avoid a “visual lie”.



Robbins, N. B. (2005)¹⁶

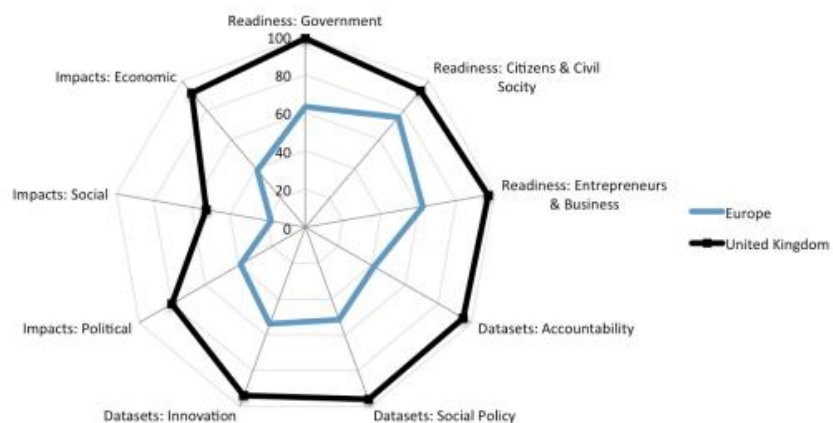
¹⁶ Robbins, N. B. (2005). *Creating more effective graphs*. Hoboken, NJ: Wiley-Interscience.

Isotype, in a simple application, are bar charts, where the bar has been replaced by symbols, e.g. pictograms for people.



Neurath, Otto (1936)¹⁷

Radar charts are especially useful if the data is cyclical such as weekdays. Other cases have to be considered carefully because marking comparisons is often hard, they become easily cluttered, and alternatives such as a bar chart may be superior. The chart from the Open Data Barometer works well because the levels are not overlapping.



<http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf>

¹⁷ Neurath, Otto (1936). International Picture Language, London: Kegan Paul, Trench, Trubner & Co.

4.4.3 Stacked bar charts

Stacked bar charts are a generalisation of bar charts. They include one additional dimension, that is the bar is broken down for example by colour. It is again an example from the PSI Scoreboard, where seven indicators compose the overall score. The use of colour is distinct (though perhaps not optimised for colour-blind people), the overall layout clean with an alphabetical order that complements the overall bar chart above, and the additional information, displayed only by mouse-over, helps explain details of the score.

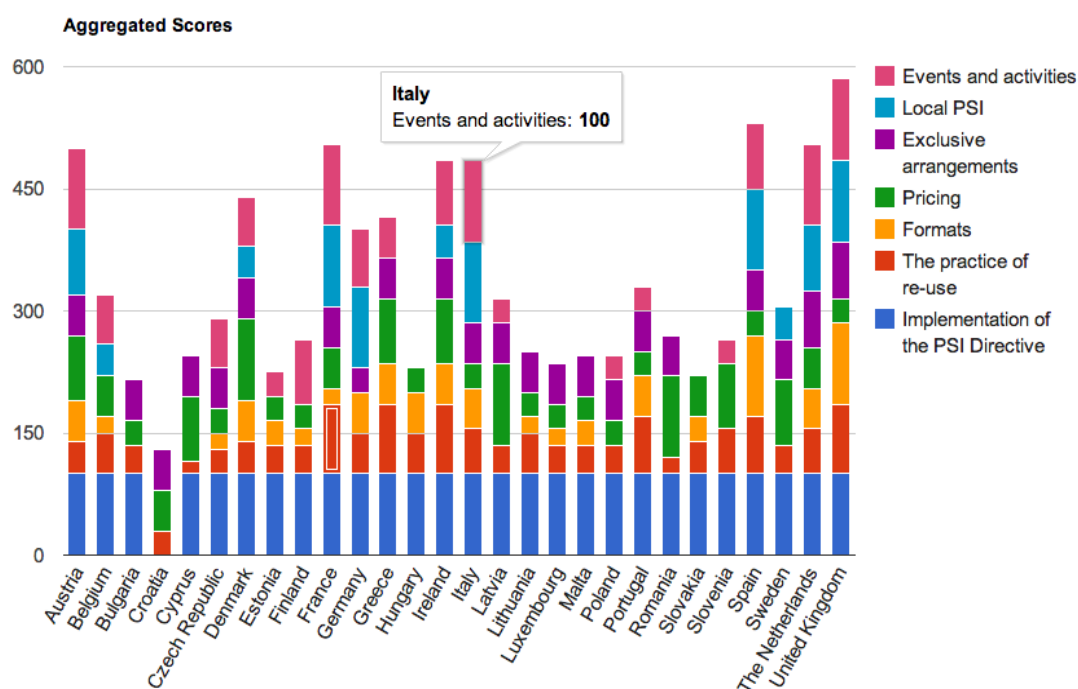


Figure 11. Example of a stacked bar chart
(Source: <http://www.epsiplatform.eu/content/european-psi-scoreboard>)

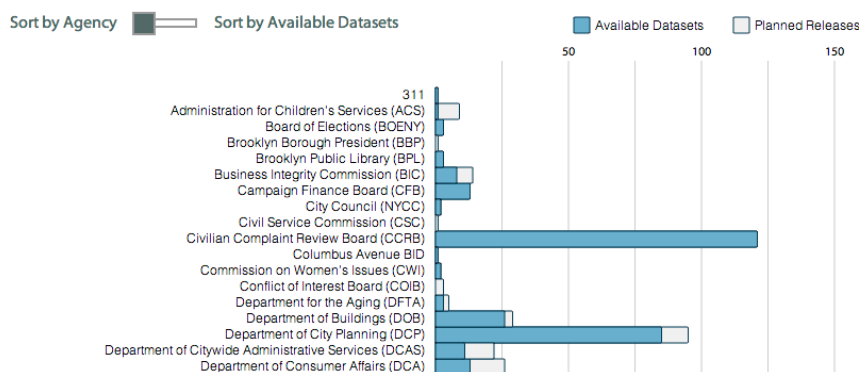
Do	Improve
Use distinctive colours, provide detailed information without being cluttered, e.g. via mouseover	Avoid too much information in one chart without guiding the user through the visualisation

Table 20. Variations of the stacked bar chart

Overlaid bars are useful if there are only a few, or arguably only two, categories. The example on the right does not take into account hidden bars, which could be solved by introducing opacity.

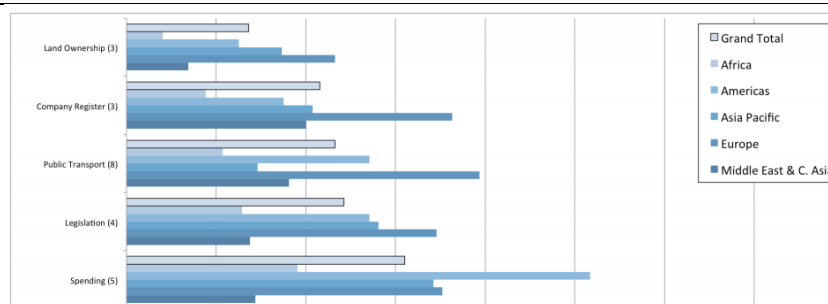
Explore Agency Status

View planned and released datasets by Agency



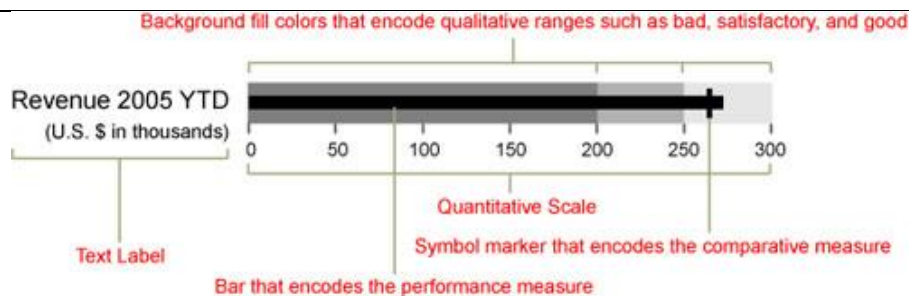
<https://data.cityofnewyork.us/dashboard>

Multiple bars also allow for an additional dimension. The example on the right could be improved by colours that are easier to distinguish.



<http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf>

Bullet graphs are useful for displaying the progress towards a target. They are often found in the context of business and dashboards.



Few, Stephen (2013)¹⁸

¹⁸ Few, Stephen (2013). Information Dashboard Design: Displaying data for at-a-glance monitoring, Second Edition, Analytics Press.

4.4.4 Line charts

A line chart is a simple visualisation technique and often very useful to display changes over time. The following example is only a mock visualisation from Veljković et al. (2014), but nonetheless a good example to emphasise a few visualisation principles.

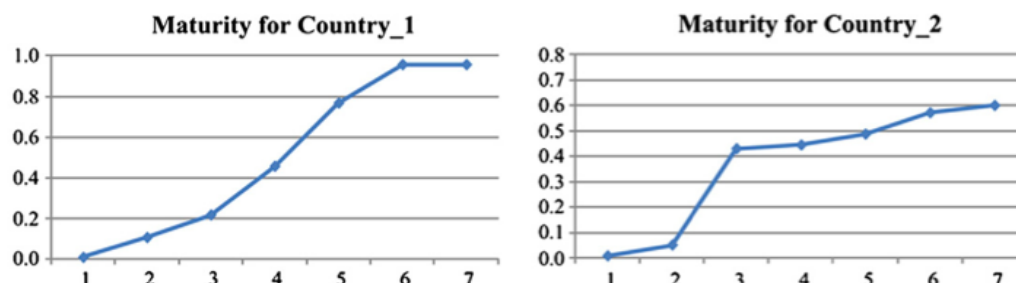
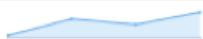







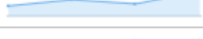

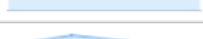
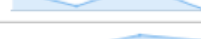

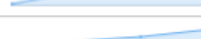








Figure 12. An example of Maturity progression. Veljković et al. (2014).

Do	Improve
Use a line chart for changes over time	Avoid different scales that make comparisons more difficult

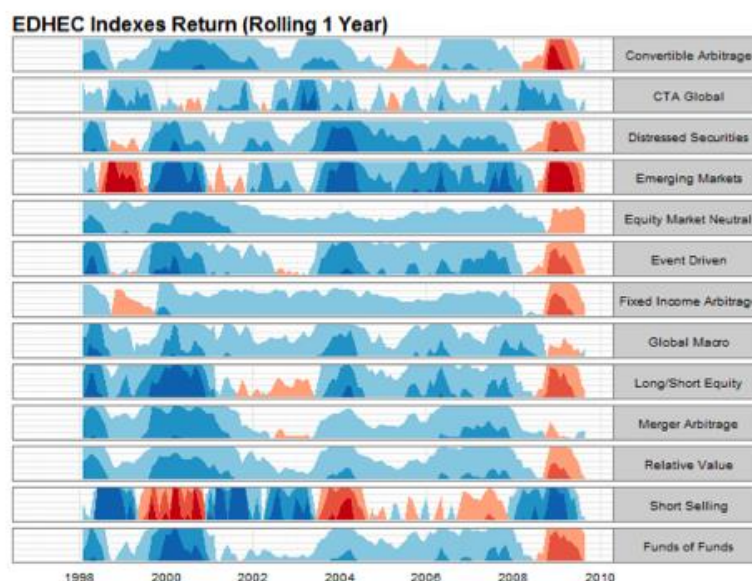
Commendable is the repeated use of the same chart also known as *small multiples* (more in section 4.4.10). However, using different vertical scales defies the purpose of making easier comparisons. The grid and axes lines are grey, which makes them a lot less intrusive, but could still be optimised, e.g. by omitting some or making them thinner.

Table 21. Variations of the line chart

Area chart	State	Income	Income per quarter	Costs	Costs per quarter	Result
	Austria	201		148		53
	Belgium	249		244		5
	Bulgaria	183		212		-29
	Croatia	232		172		60
	Cyprus	166		152		14
	Czech Republic	336		151		185
	Denmark	216		216		0
	Estonia	135		159		-24
	Finland	184		215		-31
	France	289		219		70

Own example of small area charts with mockup data.

Horizon plot



<http://timelyportfolio.blogspot.co.uk/2012/08/horizon-on-ggplot2.html>

4.4.5 Distributions and histograms

For visualising the distribution of a variable, e.g. the size of datasets in a catalogue, we require a **histogram** (or its continuous equivalent, a density estimate). Histograms are very useful to see underlying patterns in data that an average would not be able to do justice.

We find an example in the Metadata Census: below is a histogram of the quality metric “completeness”. The graph is clean and shows that the data exhibits a small bump around 40, which we would otherwise not appreciate.

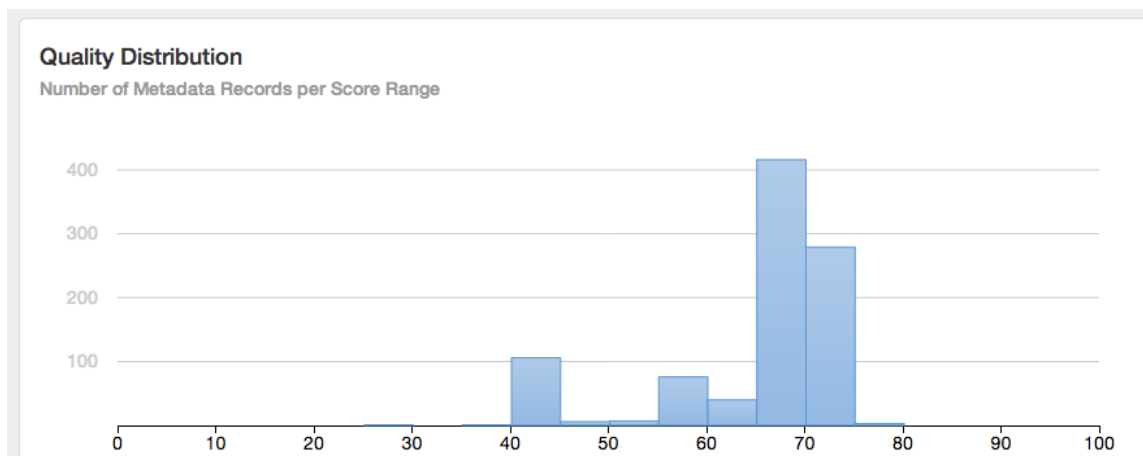


Figure 13. Example histogram

(Source: <http://metadata-census.com/repositories/data.gov.uk/snapshots/2013-11-16/metrics/completeness>)

Do	Improve
Summarise data via a histogram and provide more information than an average	Include the explanation of what is a distribution and/or histogram

4.4.6 Pie charts

The pie chart is a circular chart with sectors proportional to data. It is most useful when showing the composition of a metric, for example, relative metrics that add up to 100%. The example by the ENGAGE project shows the different licences that are used in European data portals. Again we encounter a few issues that probably stem from the automated generation of the chart:

- Labels are truncated and are therefore not readable.
- The chart displays too many data points, which defies it's purpose. One solution could be to group small categories into "other".
- The main category "license not specified" is somehow puzzling without a further explanation.
- The data label 81.74672...% suffers from pseudo-precision.

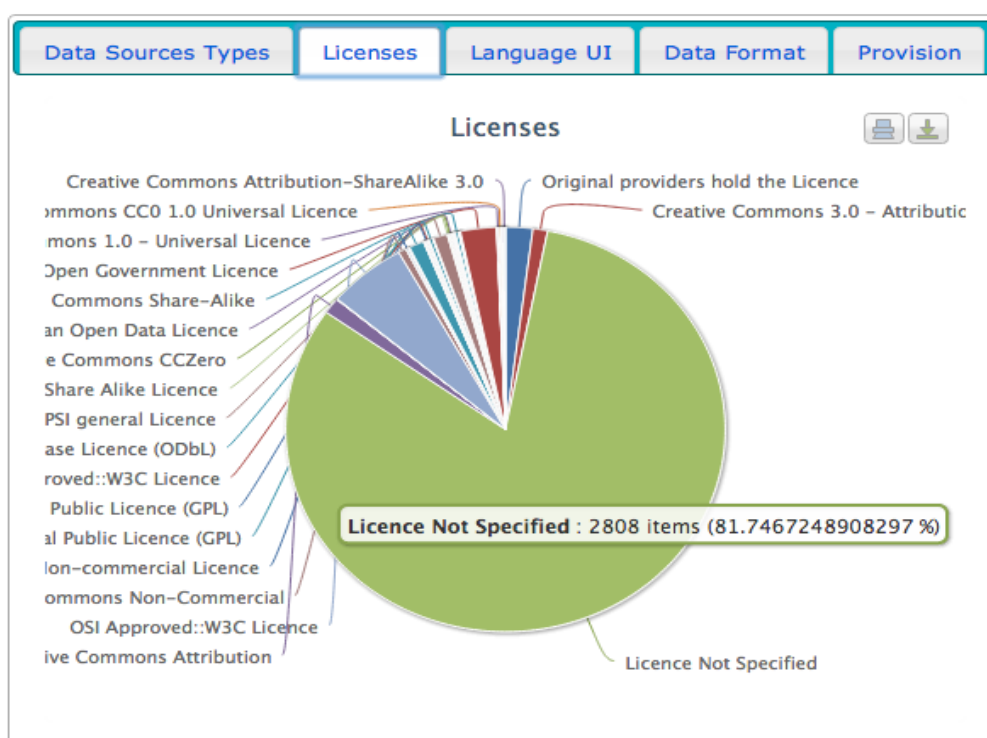
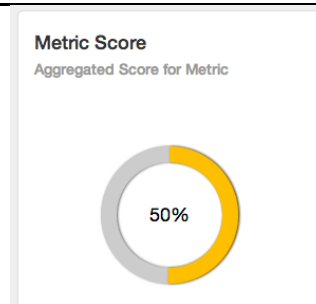


Figure 14. Example pie chart (Source: <http://www.engagedata.eu/opendatasites>)

Do	Improve
Use a pie chart if categories sum up to 100%	Avoid too many categories and numbers that are too precise

Table 22. Variations of the pie chart

The shape of the pie is sometimes a **doughnut** or **semi-circle** as shown in the example on the right.



<http://metadata-census.com/repositories/data.gov.uk/snapshots/2013-11-16/metrics/completeness>

4.4.7 Network visualisations

The open data ecosystem contains networks, so an application of this technique may be relevant to the ODM project. For example, the open datasets from New York City are linked in various ways and can be explored in an interactive version (see below). There are plenty of ways to visualise a network, although creating network visualisations often requires specialised knowledge.

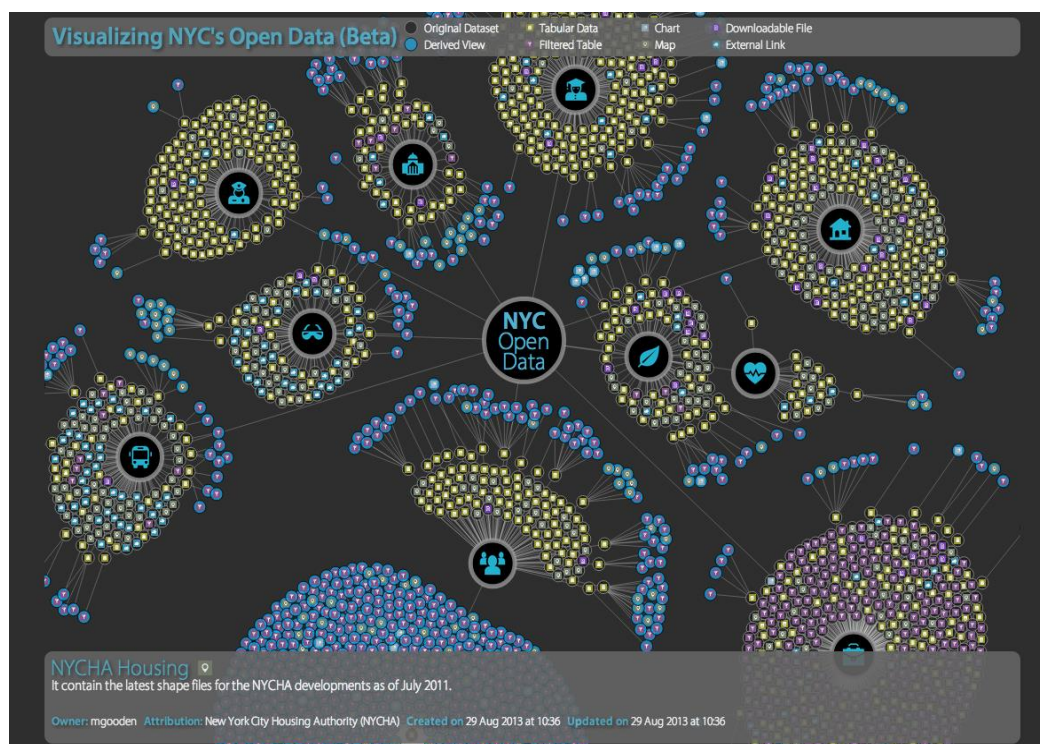


Figure 15. Example network visualisation (Source: <https://data.cityofnewyork.us/viz>)

Do	Improve
Recognise that networks require a different type of visualisation	Avoid specialised techniques where a simple chart would suffice

4.4.8 Heat maps

Heat maps are visualisations where tabular data is represented by colours.¹⁹ They are great for large matrices and as an alternative to tables. For example, the Open Knowledge Foundation uses a heat map to display discrete data (with four categories) about open data in various countries.

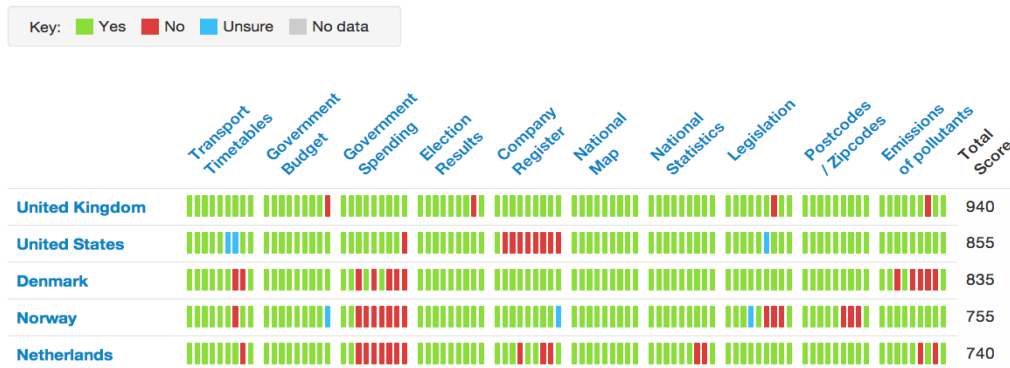


Figure 16. Example heat map (Source: <https://index.okfn.org/country>)

Do	Improve
Use colours to visualise complex tables	Remember that every choice, e.g. groupings, equal sizes, carries an implicit interpretation

A heat map can also be used with continuous scales, where the level is mapped to a colour code. We find a mix of both techniques in a visualisation of various metadata scores below. The table is turned into a heat map, where different levels are highlighted by a discrete colour scale. (The labels on top could be improved by including line breaks instead of rotating them.)

Rank	Repository	IF Score	Completeness	Weighted Completeness	Openness	Accuracy	Richness Of Information	Intrinsic Precision	Readability	Availability
1	data.sa.gov.au	0.81	0.73	0.86	1.00	0.84	0.53	0.98	0.73	0.84
2	africaopendata.org	0.70	0.64	0.59	0.75	0.87	0.36	1.00	0.54	0.86
3	data.qld.gov.au	0.66	0.69	0.72	0.96	0.41	0.31	0.99	0.59	0.64
4	datos.codeandmexico.org	0.61	0.55	0.60	0.64	0.00	-	1.00	0.45	1.00
5	GovData.de	0.60	0.55	0.61	0.44	0.02	0.44	0.99	0.79	0.94
6	data.gv.at	0.57	0.50	0.48	0.99	0.00	0.31	1.00	0.58	0.72
7	dados.gov.br	0.57	0.53	0.59	0.37	0.28	0.77	0.96	0.45	0.62
8	catalogodatos.gub.uy	0.55	0.64	0.70	0.01	0.14	0.52	0.97	0.65	0.77
9	data.openpolice.ru	0.53	0.56	0.63	0.00	0.00	0.09	1.00	1.00	1.00
10	data.gov.uk	0.51	0.53	0.55	0.53	0.21	0.44	0.99	0.54	0.26
11	data.gc.ca	0.46	0.53	0.56	0.00	-	-	-	-	0.74
12	opendata.admin.ch	0.44	0.52	0.58	0.00	0.00	0.31	1.00	0.35	0.77
13	data.gov.sk	0.39	0.45	0.45	0.00	0.00	0.57	1.00	0.37	0.25

Figure 17. Example tabular heat map with continuous scales
(Source: <http://metadata-census.com/repositories#leaderboard>)

¹⁹ Sometimes the term is also used for (continuous) choropleth maps.

Do

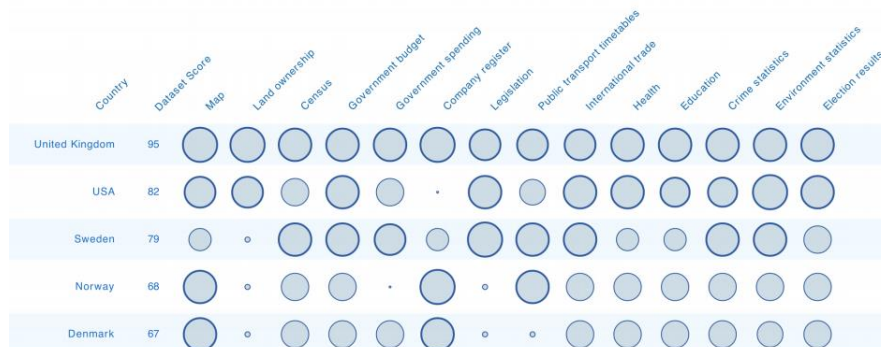
Use conditional formatting such as font colours even for a table

Improve

Do not include more than two or three significant digits

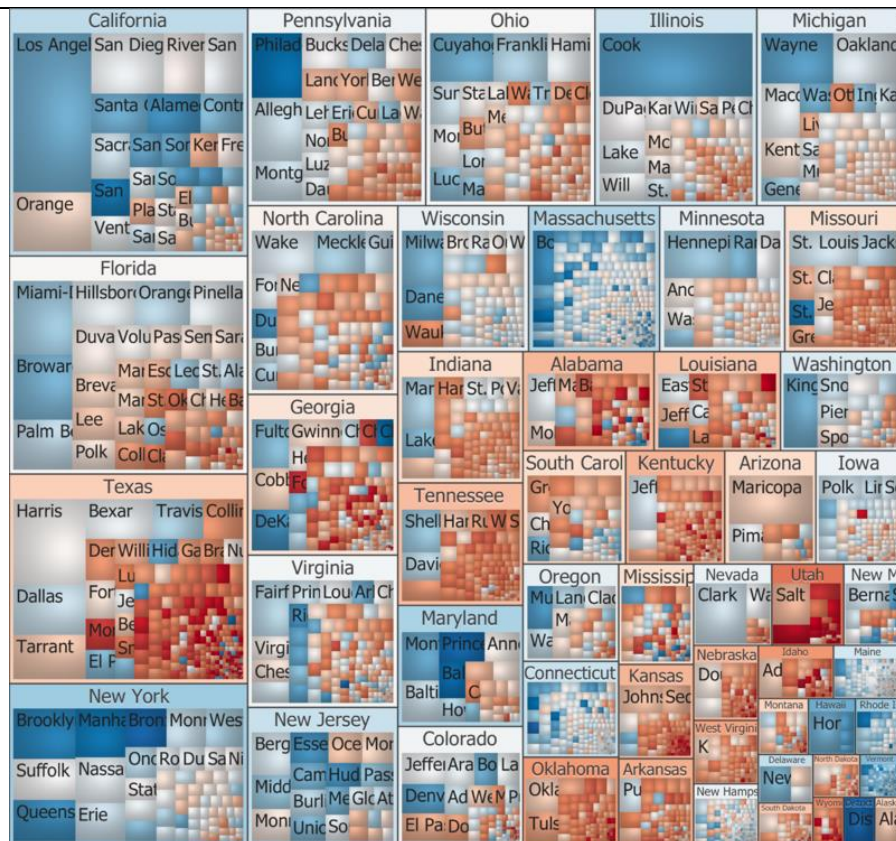
Figure 18. Variations of the heat map

A **symbolic heat map** from the Open Data Barometer, where different, qualitative levels of openness are represented by symbols.



<http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf>

A **tree map** of votes by county, state and locally predominant recipient in the US Presidential Elections of 2012.



Source: Luc Girardin

http://commons.wikimedia.org/wiki/File:US_Presidential_Elections_2012.png#mediaviewer/File:US_Presidential_Elections_2012.png

4.4.9 Scatter plots

Scatter plots, and its extensions such as bubble charts, are ideal to visualise relationships between metrics. One of the most recognisable uses of this technique are visualisations of open data published by the World Bank. Hans Rosling and his team have brought it, via the tool Gapminder, into the mainstream. The screenshot below shows the relationship between internet users and income and includes further dimensions such as time (dynamic, on the bottom), population (bubble size) and geography (colour).

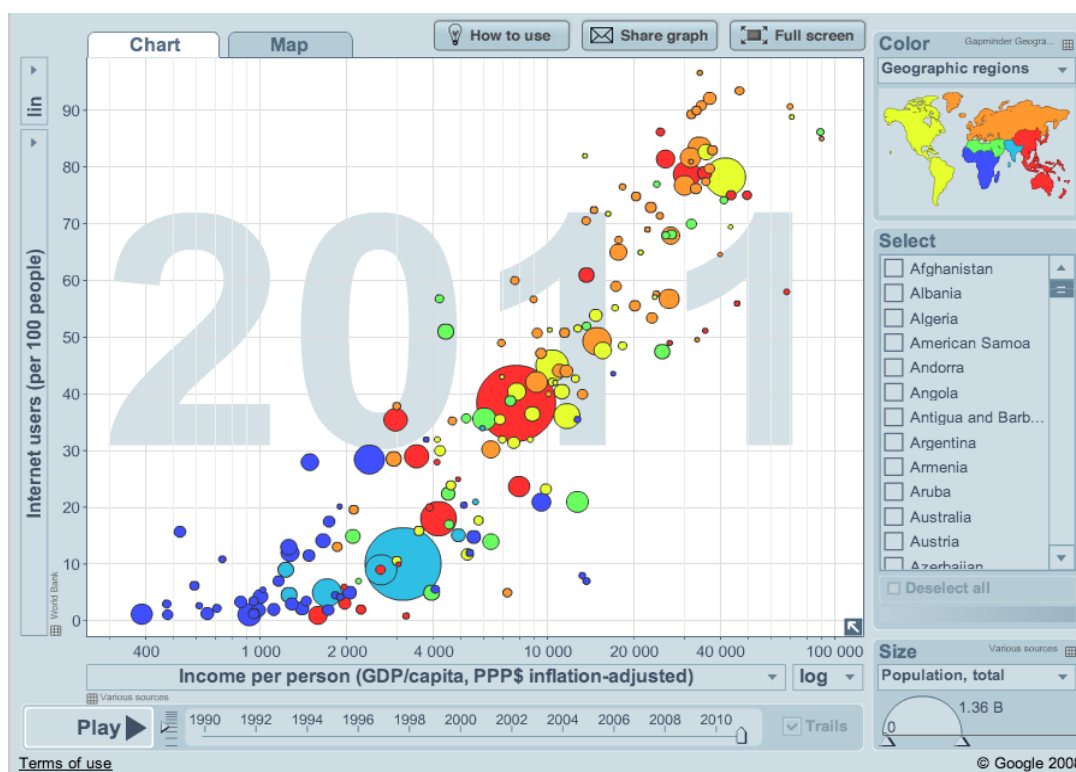


Figure 19. Example scatter plot (Source: <http://www.gapminder.org/>)

Do

Use size, shape, colour, animation etc to visualise additional dimensions

Improve

Avoid correlations that show a spurious causal relationship

The ODM project, with its descriptive nature, contains few relationships of continuous variable. Thus, scatter plots are mostly out of scope.

4.4.10 Meta-techniques: tables, sparklines, small multiples, composite graphs

There are some visualisation techniques that we consider a “meta-technique” as they are not exclusive to one type of chart. They span most visualisations because they are considerations on how to display data regardless of whether it is a bar or a line chart.

Tables

A table is not a visualisation, however, it is sometimes a lot more useful than an complicated stacked bar chart. It is up to the researcher to decide in what instances a table might be a viable, or better, alternative. Again it relates to considerations around context and audience. A table is particularly useful if the numbers themselves, and not just the comparison among them, are relevant to the interpretation.

Sparklines

Edward Tufte’s sparklines are “data-intense, design-simple, word-sized graphics”. They are small, possibly in-line, visualisation that carry a high density. A playful application of this principle is the following tweet from the Wall Street Journal.²⁰



Figure 20. Example spark line

Small multiples

Some charts can be repeated for each category: for example, a line chart that shows the count of data catalogues for each country over time. If one chart is explained, by including axes and axes labels and further annotations, the reader can easily infer how to interpret all other charts. Hence, the idea of multiples speeds up understanding and consumption of many charts. This works even when they are small and are missing some annotations. Crucial, however, is to repeat the chart format and, for example, keep the limits of the axes the same.

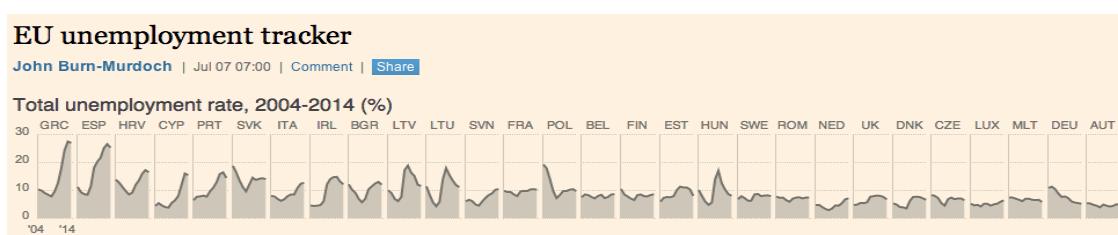


Figure 21. Example of small multiples

(Source: <http://blogs.ft.com/ftdata/2014/07/07/eu-unemployment-tracker>)

Composite graphs

Combining different charts to a more comprehensive view often adds value beyond the sum of its parts. As mentioned before, contextual information brings meaning to a visualisation and several charts next to each other may serve this purpose. Sometimes several charts provide an opportunity to address several audiences: those who want a quick overview, and others that may be more

²⁰ <https://twitter.com/WSJ/status/66484941051019265>

interested in detailed information. The lines on when it becomes a dashboard are blurred; more on how to design an exemplar dashboard is in the section on dashboards.

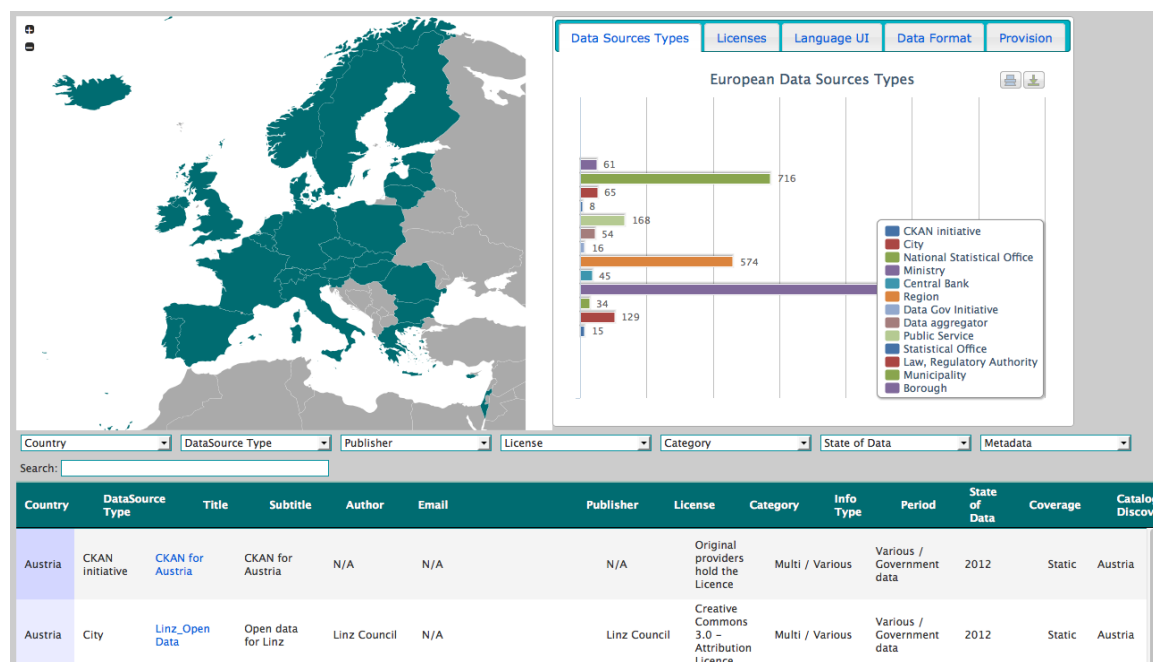


Figure 22. Example composite graph
(Source: <http://www.engagedata.eu/opendatasites>)

Do	Improve
Join various visualisation methods and, if useful, include the underlying data	Target the right audience and do not forget their putative data literacy

4.4.11 Bespoke infographics

The optimal, or minimal, graphs are seldom novel and there is an argument that an exemplar presentation of data also *engages* the user. Many designers are therefore willing to trade off some of the functional aspects of a visualisation with stylistic choices that raise the interest of an audience. However, the expertise and balance to achieve this is extremely delicate and often falls short of the expectations. The ODM project will mostly focus on functional graphs.

Below we see an example commissioned by the OECD. It is essentially a variation of the radar chart in a bespoke and interactive application. Precise metrics and comparability are not easy to gauge but the visualisation excels in a novel, and arguably engaging, comparison of the Better Life Index across countries.

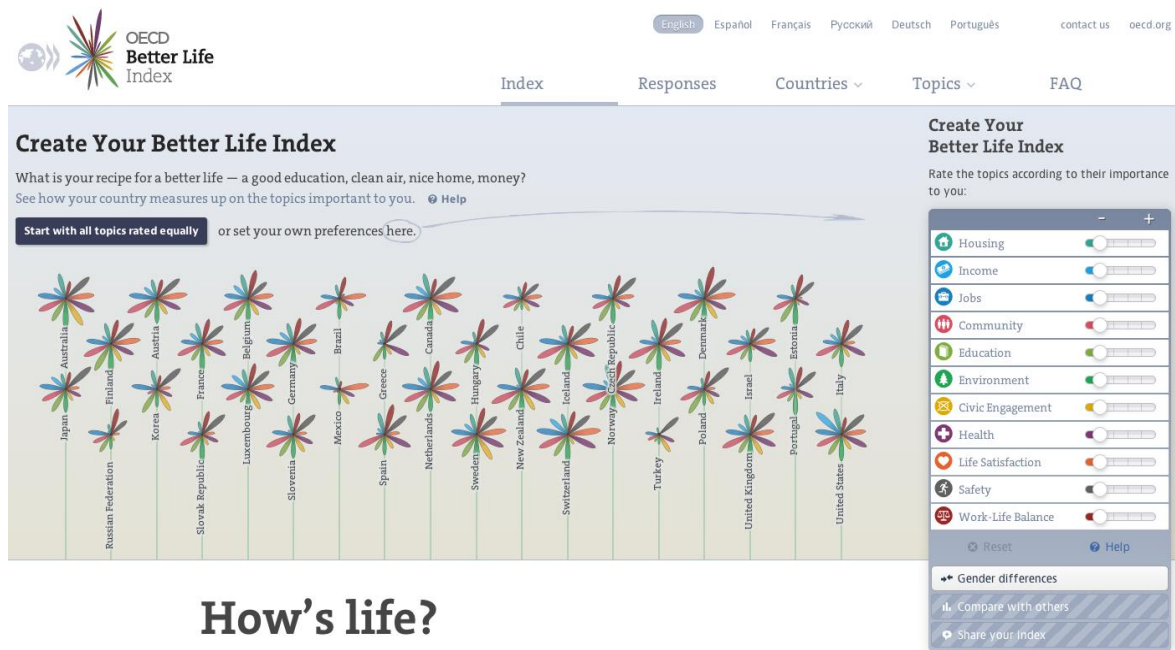


Figure 23. Example bespoke infographic
(Source: <http://www.oecdbetterlifeindex.org>)

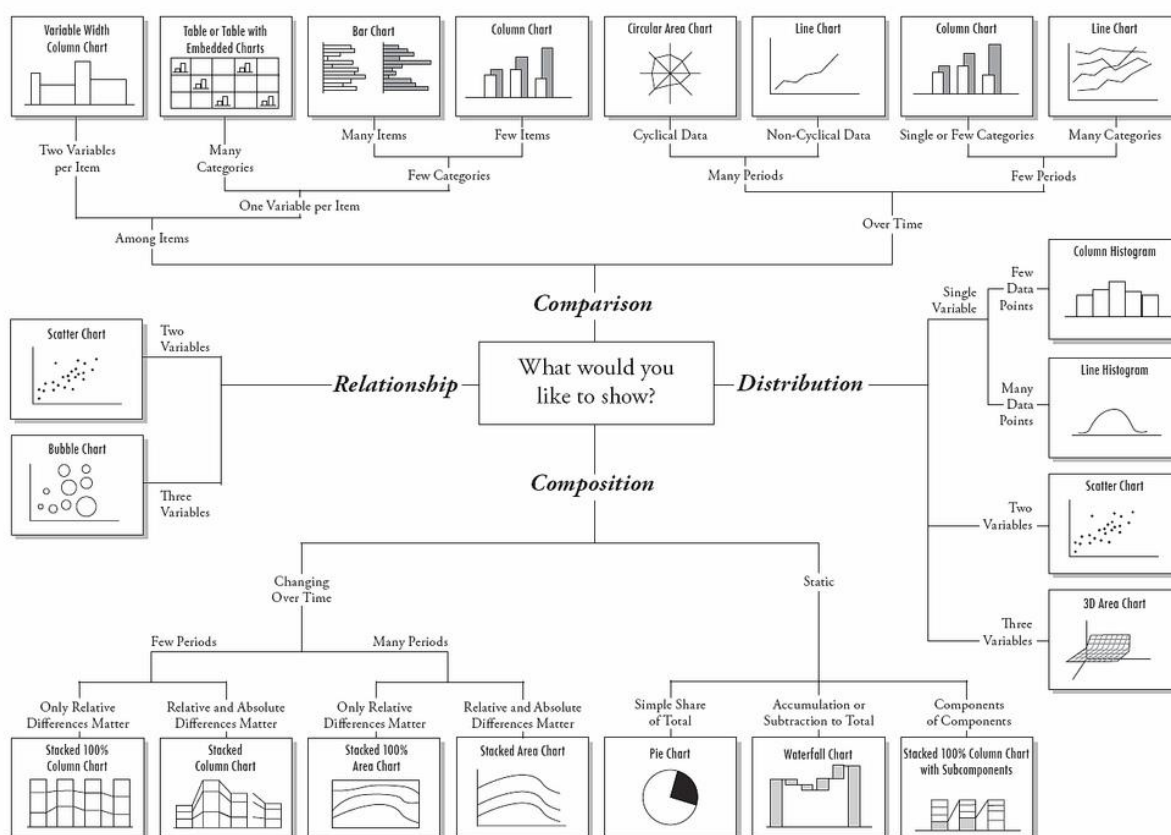
Do	Improve
Engage the user with a novel and/or interactive visualisation	Avoid an infographic for the sake of it

4.5 Selection of visualisation techniques

As we have shown, the menu and design choices for visualising data are enormous. Our selection is motivated by a combination of reasons:

- Some techniques suggest themselves naturally, for example a line chart for a time series. See also Figure 24. Chart Chooser for a starting point.
- In other cases we stick to visualisations that are most readily understood by humans such as comparisons on a common scale, which for example leads to bar charts.
- Variety in the techniques, if used with caution, can make visualisation more engaging.
- We prefer common techniques, i.e. those that are familiar to most people and hence need not be explained in detail.
- Another factor is the technical difficulty of implementing a visualisation technique: we give preference to simpler ones.

Chart Suggestions—A Thought-Starter



© 2006 A. Abela — a.vabela@gmail.com

Figure 24. Chart Chooser

Source: Andrew Abela, <http://extremepresentation.typepad.com/blog/2008/06/visualization-taxonomies.html>

Table 23. Overview of ODM's main visualisation techniques

Type	Use	Relevance to ODM
Figure/table	Not a visualisation as such, but often a useful choice for a small number of data points or when the figure itself is important.	Aggregate scores, high-level counts, rankings
Bar chart	Ideal choice for comparisons, e.g. for a metric with many categories.	Comparison of data formats, licences etc.
Line chart	Ideal choice for comparisons over time.	Metrics that are displayed with the context of time, e.g. increase of number of data catalogues over the years.
Pie chart	Ideal choice for a composition	Examples include sectors in a catalogue, languages etc.
Graduated	Comparisons across countries that	Any score on a country, or regional, level.

symbol map	avoids the area bias	
Heat map	Ideal choice for tables; can be qualitative or quantitative	Complex rank tables, e.g. visualising several scores across countries at once.
Histogram	Ideal choice for distributions	Size of datasets within a catalogue etc.
Other variations	E.g. dot plots, multiple bar charts	Where appropriate
Meta-techniques	Sparklines, small multiples, composite graphs	For dashboard views

4.6 Mapping of metrics to visualisation techniques

4.6.1 Measures over the aggregate

<i>Counts/ Averages/ Longitudinal</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/ heat map	Other	Meta-techniques
Total number of catalogues	x							
Frequency of catalogues by sector of publishing organisation		x						
Proportion of catalogues by sector of publishing organisation				x				
Frequency of catalogues using specific software platforms		x						
Proportion of catalogues using specific software platforms				x				
Median age of catalogues	x							
Mean age of catalogues	x							
New catalogues per month			x					

<i>Rankings</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/ heat map	Other	Meta- techniques
Highest frequency of catalogues per country						x		
Lowest frequency of catalogues per country						x		
Highest frequency of catalogues per capita per country						x		
Lowest frequency of catalogues per capita per country						x		

4.6.2 Per-geography measures

<i>Catalogue statistics</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/ heat map	Other	Meta- techniques
Catalogues per geographic region		x			(x)			
Catalogues per capita per country		x			(x)			
Catalogues & per-capita GDP correlation (Pearson and Spearman's rank)					(x)	x		
Catalogues & HDI correlation (Pearson and Spearman's rank)					(x)	x		
Frequency of catalogues by sector of publishing organisation		x						
Proportion of catalogues by sector of publishing organisation				x				

<i>Time-profile by country</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
New catalogues per country per month			x					Small multiples

4.6.3 Per-catalogue measures

<i>Data Volumes</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Frequency of catalogued datasets	x	(x)						combine with resources
Frequency of catalogued distributions	x	(x)						combine with collections
Frequency of unique organisations publishing data		x						
Total distribution size in a catalogue	x						(histo gram)	
Median distribution size	x						(histo gram)	
Mean distribution size	x						(histo gram)	
Maximum distribution size	x						(histo gram)	
Standard deviation of distribution sizes	x						(histo gram)	

<i>Data Duplication/Uniqueness</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Proportion of distributions in each catalogue that are <i>listed</i> in other catalogues	x							
Proportion of distributions in each catalogue that are <i>not listed</i> in any other catalogues	x							

<i>Housekeeping</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Proportion of data file links that are broken				x				
Proportion of different HTTP status codes for data file URIs				x				

<i>Formats and machine-readability</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Frequency of distributions by file format		x						
Proportion of distributions by file format				x				
Frequency of distributions in a machine-readable file format		x						
Proportion of distributions in a machine-readable file format				x				
Frequency of distributions by MIME type of data file		x						
Proportion of distributions by MIME type of data file				x				
Frequency of distributions that are machine-readable		x						
Proportion of distributions that are machine-readable				x				

<i>Licenses</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Frequency of distributions with an explicitly set license	x							
Proportion of distributions with an explicitly set license							bullet chart	

Frequency of distributions with an open license	x							
Proportion of distributions with an open license (excluding and including datasets with missing licenses)							bullet chart	
Frequency of distributions by license type		x						
Proportion of distributions by license type (excluding and including datasets with missing licenses)				x				

<i>Release Frequencies and Timeliness</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Median days since latest dataset update	x							
Median days since latest new dataset	x							
Frequency of dataset last update by year		x						
Frequency of datasets with stated update frequency		x						
Proportion of datasets with stated update frequency				x				
Tau of the catalogue							bullet chart	

<i>Prominence, Engagement and Usability</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
PageRank of the catalogue site	x							
Frequency of unique publishers contributing to the catalogues						x		

Frequency of unique publishers relative to catalogue size						x		
Frequency of datasets available via APIs and/or data dumps		x						
Proportion of datasets available via APIs and/or data dumps				x				
Ratio of datasets with APIs to those with data dumps						x		
Frequency of distributions with previews		x						
Proportion of distributions with previews				x				
Frequency of different languages		x						

4.6.4 Per-dataset measures

<i>Data and Metadata Volume, Quality and Usability</i>	Figure /table	Bar chart	Line chart	Pie chart	Map	Ranking/heat map	Other	Meta-techniques
Dataset size	x							(x)
Number of fields in the metadata record that are populated	x							(x)
Frequency of unique vocabularies used in metadata record	x							(x)
Frequency of terms used from each vocabulary present in metadata record		x						(x)
Proportion of terms used from each vocabulary present in metadata record				x				(x)
Open Data Certificate level of the dataset							badge	

Frequency of Errors and Warnings generated by CSVlint							badge	
Timeliness of the dataset							badge	

4.7 Software Libraries for Visualisation

As already mentioned, modern computer technology can be used for creating visualizations of any kind in a convenient way practically without limits. Also in the web domain, as major technologies have grown mature and myriads of toolkits and web frameworks have evolved making life easier for web platform developers, libraries and toolkits for visualization issues have been created. In this section an overview about the general availability shall be provided. For a detailed analysis about the applicability of the solutions for ODM please refer to deliverable D3.2.

4.7.1 Libraries Overview

These days, web pages are usually designed in an eye-catching, responsive and intuitive manner with the aid of technologies like Flash or JavaScript. Visualization libraries fit well into that picture by transforming raw data into visually attractive charts created dynamically at runtime. The most common technology used therefore is JavaScript, which further holds the advantage of being free, open source and platform independent. Combined together with other open standards such as HTML, CSS, DOM, XML and JSON, a whole world of libraries and frameworks opens up, offering services on different levels, from very low-level plotting facilities that offer a huge flexibility, to high-level chart types that just need to be provided with the actual data and a few parameters.

Here, the focus is put on tools that are ready-to-use and provide high-level chart types like bar, line or pie charts out of the box. Many of them offer several chart types at once, but there are also some specialized on specific ones like overlay maps (choropleth map) or time series charts for instance. Another discriminatory factor among them is whether charts produced are interactive, meaning that the user can actually interact with the charts like for instance change the sort key or arrangement of elements, or whether the result is rendered as scalable vector graphic (SVG), vector markup language (VML) for legacy support, or an HTML5 canvas. From the development point of view the grade of customizability, easiness to learn and of course the license represent the most important factors. As browser compatibility problems have decreased over the past years, this is not an issue for the most chart libraries either, as they are even supported on most mobile device browsers.

The following Table 24 shows an overview about the solutions coming into question for our purpose. It contains only libraries that support at least bar charts, line charts and pie charts at once or time line charts or choropleth maps. The list surely is not exhaustive but covers the most common ones. Further, only those solutions were considered that are either open source or at least free to use for a non-profit platform such as ODM.

Table 24. Overview of all JavaScript visualization libraries and frameworks that are open source or at least free for a NPO.

Name	URL	License	Bar/Line/Pie Chart	Time Line Chart	Choropleth Map	Interactivity	HTML5 Canvas	SVG	Comment
AwesomeChart JS	cyberpython.github.io/AwesomeChartJS	Apache 2.0	x				x		
CanvasJS	canvasjs.com	CC BY-NC 3.0	x	x		x	x		
canvasXpress	canvasxpress.org	GPL 3.0	x			x	x		
ccchart	ccchart.com	MIT	x				x		
Chart.js	www.chartjs.org	MIT	x			x	x		
Chartkick	ankane.github.io/chartkick	MIT	x		x	x		x	
Cubism.js	square.github.io/cubism	Apache 2.0		x			x		Based on D3.js
D3.js/Protovis	d3js.org	BSD	x	x	x	x		x	
DataMaps	datamaps.github.io	MIT			x	x		x	Based on D3.js
dc.js	nickqizhu.github.io/dc.js	Apache 2.0	x	x		x		x	Based on D3.js
dhtmlxChart	www.dhtmlx.com/docs/products/dhtmlxChart/index.shtml	GPL	x			x	x		
Dojo Charting	dojotoolkit.org/features/graphics-and-charting	BSD	x		x	x	x	x	
dygraphs	dygraphs.com	MIT		x		x	x		
Elycharts	elycharts.com	MIT	x			x		x	Based on Raphaël
EmberCharts	addepar.github.io/#/ember-charts/overview	Apache 2.0	x	x		x			Based on D3.js
Envision	www.humblesoftware.com/envision	MIT		x		x	x		
Flot Charts	www.flotcharts.org	MIT	x			x	x		
Flotr2	www.humblesoftware.com/flotr2	MIT	x			x	x		

Google Charts	developers.google.com/chart	Apache 2.0	x	x	x	x		x	
gRaphaël	g.raphaeljs.com	MIT	x					x	
Highcharts	www.highcharts.com	Free for NPO	x	x	x	x		x	
jqPlot	www.jqplot.com	MIT/GPL 2.0	x				x		
jVectorMap	jvectormap.com	Free			x	x		x	
Kartograph	kartograph.org	LGPL			x	x		x	Based on Raphaël
Leaflet	leafletjs.com	BSD			x	x	x		
MilkChart	mootools.net/forged/p/milkchart	MIT	x				x		
morris.js	morrisjs.github.io/morris.js	BSD	x			x		x	Based on Raphaël
NVD3	nvd3.org	Apache 2.0	x			x		x	Based on D3.js
OLAPCharts	www.olapcharts.com	Free	x			x	x		
PlotKit	www.liquidx.net/plotkit	BSD	x				x	x	
Polymaps	polymaps.org	Free			x			x	
RGraph	www.rgraph.net	MIT	x			x	x		
Rickshaw	code.shutterstock.com/rickshaw	MIT		x		x			Based on D3.js
Shield UI	www.shieldui.com	Free for NPO	x			x		x	
TimeChart	timechart.toolset.io	Free for NPO		x		x	x		
Timeplot	www.simile-widgets.org/timeplot	BSD		x		x	x		
Vega	trifacta.github.io/vega	Free	x		x		x	x	Based on D3.js
ZoomCharts	zoomcharts.com	Free for NPO	x	x	x	x	x		

4.7.2 Common Solutions Analysis

The most popular solutions were identified by conducting a Google Trend analysis. The results are shown in Figure 25. According to it, D3.js, Google Charts and Highcharts are most popular today. Less popular but still more common compared to the remaining solutions are jqPlot, NVD3 and Chart.js.

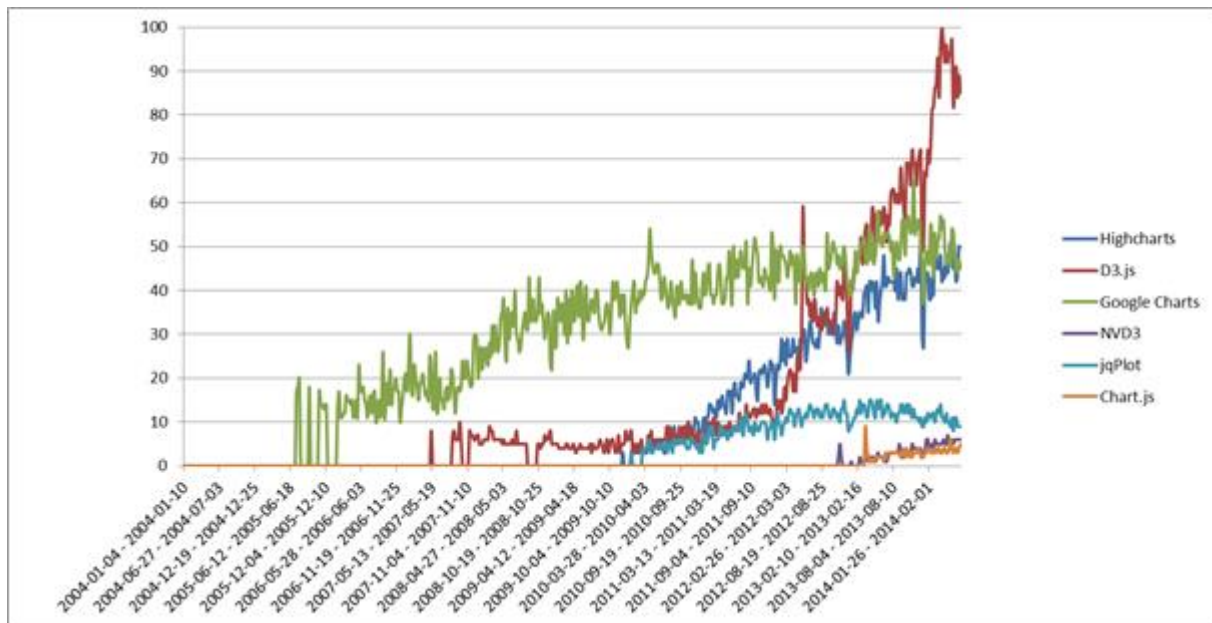


Figure 25. Google Trend analysis of the most common charting libraries of the past 10 years

In the following sections each of these common solutions is briefly examined in order to convey the range of possibilities and point out some particular features. The libraries are ordered according to their popularity.

4.7.2.1 D3.js

D3.js is not a toolkit but rather a framework (or visualization kernel) as it provides a very high flexibility and low-level function range. As a consequence, its learning curve is very long which is why many different libraries exist that built on top of it acting as a wrapper. The showcases presented on the framework's homepage, as shown in Figure 26, provide a glimpse about the vast possibilities of the tool. In general, it is best when used for novel and highly interactive charts.



Figure 26. An excerpt of the possibilities that D3.js provides as shown on the solution's home page

4.7.2.2 Google Charts

The visualization library provided by Google represents the basis of its own internal products such as Google Analytics. Unlike D3.js which has gained most of its popularity in the last two years, Google Charts has been very popular since its initial release. As a comprehensive high-level library all major chart types are supported including time series charts and choropleth maps. A selection of them is shown in Figure 27. Charts can be created rather easily, providing a solid interactivity, as indicated in Figure 28.

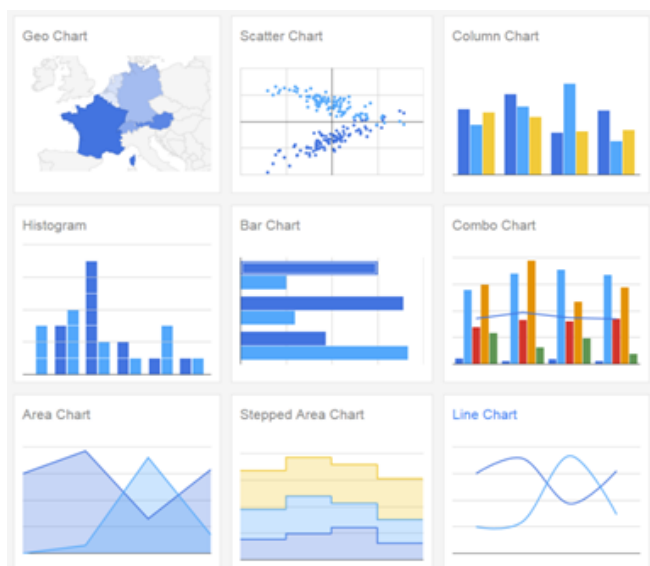


Figure 27. An excerpt of the chart types

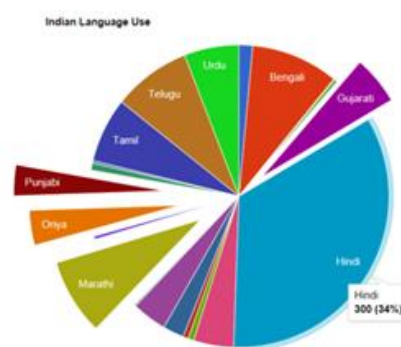


Figure 28. Interactivity features

offered by Google Charts

4.7.2.3 Highcharts

Highcharts is a highly praised library that offers a solid range of functionality and chart types, including maps and time line series, just like D3.js and Google Charts. Despite being proprietary, there is an option for free use for non-profit organizations. Its popularity is comparable to the one of Google Charts today. It is often considered as one of the most convenient solutions when it comes to customizability and flexibility. It also offers nice interactivity features as shown in Figure 29 by the means of a clickable map that alters the chart displayed on the right.



Figure 29. Combination of a choropleth map and a line chart using Highcharts

4.7.2.4 jqPlot

Although only standard chart types are supported, jqPlot represents a credible open source alternative to the previous libraries. Its usability is also considered good from the development point of view, but interactivity is not a central issue. Further, the popularity of the tool has started to decline in recent years.

4.7.2.5 NVD3

Of the solutions built on top of D3.js, NVD3 represents the most popular one. As such, it offers very convenient predefined kinds of charts that can be further customized by making use of D3.js code. The animations and interactivity features, as shown in Figure 30, make this library a serious competitor to other established high-level solutions.

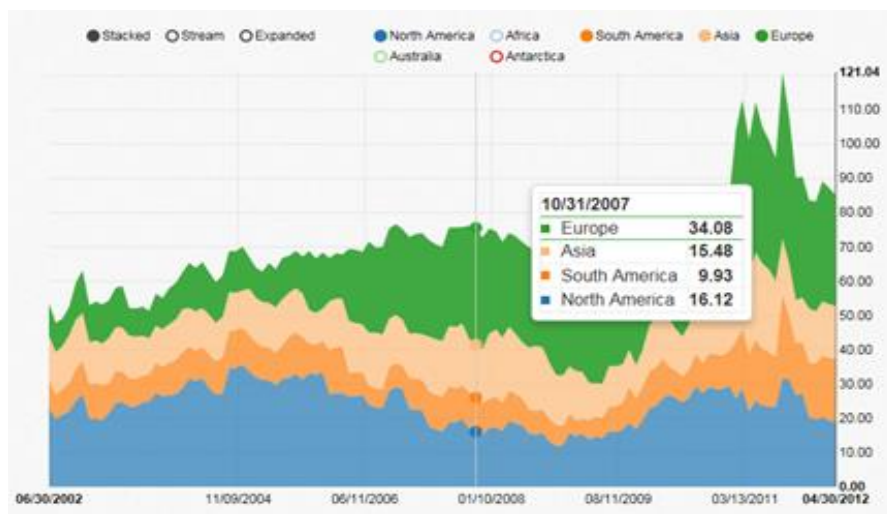


Figure 30. A highly flexible and interactive stacked area chart rendered by NVD3

4.7.2.6 Chart.js

Another reasonable open source solution is Chart.js which offers all major types of charts and is simple to use. It further offers some animation and interactivity features which cannot meet the possibilities of Highcharts or NVD3, however.

4.7.2.7 Envision

Despite not being among the top-popular list, a pure time series charting library shall be mentioned here. Envision is an open source solution which renders the resulting chart in an HTML5 canvas. The user is enabled to shift and alter the size of the visualization window by dragging the box on the bottom overall view, as shown in the small example in Figure 31.

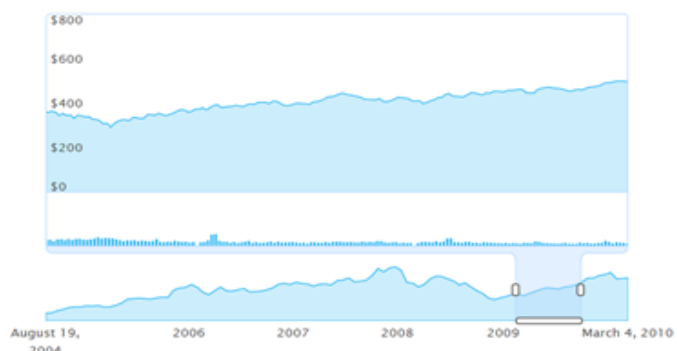


Figure 31. A time series chart where it can be altered using the overall bar on the bottom

5 Creating a dashboard

Having presented a suite of metrics and corresponding visualisation techniques that can be applied in OpenDataMonitor, this section will discuss information dashboards and explore how these may be applied in the context of ODM.

5.1 Introduction to dashboards

At its core a dashboard aids communication. The most popular form is an arrangement of critical information on a single computer screen. Stephen Few (2004), ²¹one of the leading thinkers of dashboard design, defines a dashboard as follows:

“A dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance.”

Dashboards are highly visual because graphics can communicate more efficiently than text alone. Crucial elements of the above definition are:

- its emphasis on a single screen (“information at a glance”), and;
- the context, that is the careful considerations around what are the objectives and who is the audience.

Dashboards are relatively new and their popularity have increased with the advent of IT. Software vendors have embraced it as a profitable product they can sell and there exist more and more software-as-a-service solutions. Some authors argue that the Enron scandal in 2001 put a spotlight on dashboards because managers wanted to assure shareholder they are in control. Together with key performance indicator, dashboards are a very popular tool among executive managers in all sectors.

5.1.1 Goals for an OpenDataMonitor dashboard

The dashboards are aimed at the users of the ODM: developers, entrepreneurs, civil society, policy makers, enthusiasts and others. The audience and scope is therefore not specialised or targeted at a particular group of people unlike, for example, some executive management dashboards. This implies that the dashboards are also accessible to non-technical people and do not require much time to grasp the top-level information. A menu, search function or similar navigational elements should guide the users given their needs and interests.

All outstanding dashboards are customised. In this spirit we are tailoring the ODM dashboard to our specific requirements and designing it to communicate with a broad audience. One, or several, dashboards for the ODM project aim at the following three goals.

- The dashboard **acts as a beacon** for the evolution of open data. For example, if a country or region is not keeping up with the general trend, we will discern the shortcomings timely.

²¹ Few, S. (2004). *Show me the numbers: Designing tables and graphs to enlighten* (Vol. 1, No. 1). Oakland, CA: Analytics Press.

- Displaying the most relevant information in a dashboard will function as a **reference and evidence base**. Reports, presentation or other forms of communication can refer to a simple and accessible source of information.
- We design the dashboard to **encourage better publication of open data**. This means, for example, that metrics on the dashboard are actionable.

5.2 Recommended techniques and examples

Cookbook recommendations for dashboards are very difficult because of the highly tailored nature of most designs. We will therefore illustrate a few practices by example. They mostly relate to the open data world and are relevant beyond the topic. After each example a short summary ties it back to the ODM project.

As mentioned in the introduction, dashboards are highly visual. The leading practice for visualisations of the previous section apply. Crucial elements are, moreover, the principle of “information at a glance” and the context.

It may be easier to point out a few common mistakes to avoid (adopted from Few, 2013). They are as follows:

- exceeding the boundaries of a single screen
- supplying inadequate context for the data
- displaying excessive detail or precision
- misusing or overusing colour
- introducing clutter
- choosing inappropriate or subprime visualisations (e.g. 3D-charts, gauges, radar charts)

5.2.1 Perceptual Edge Dashboard Design Competition

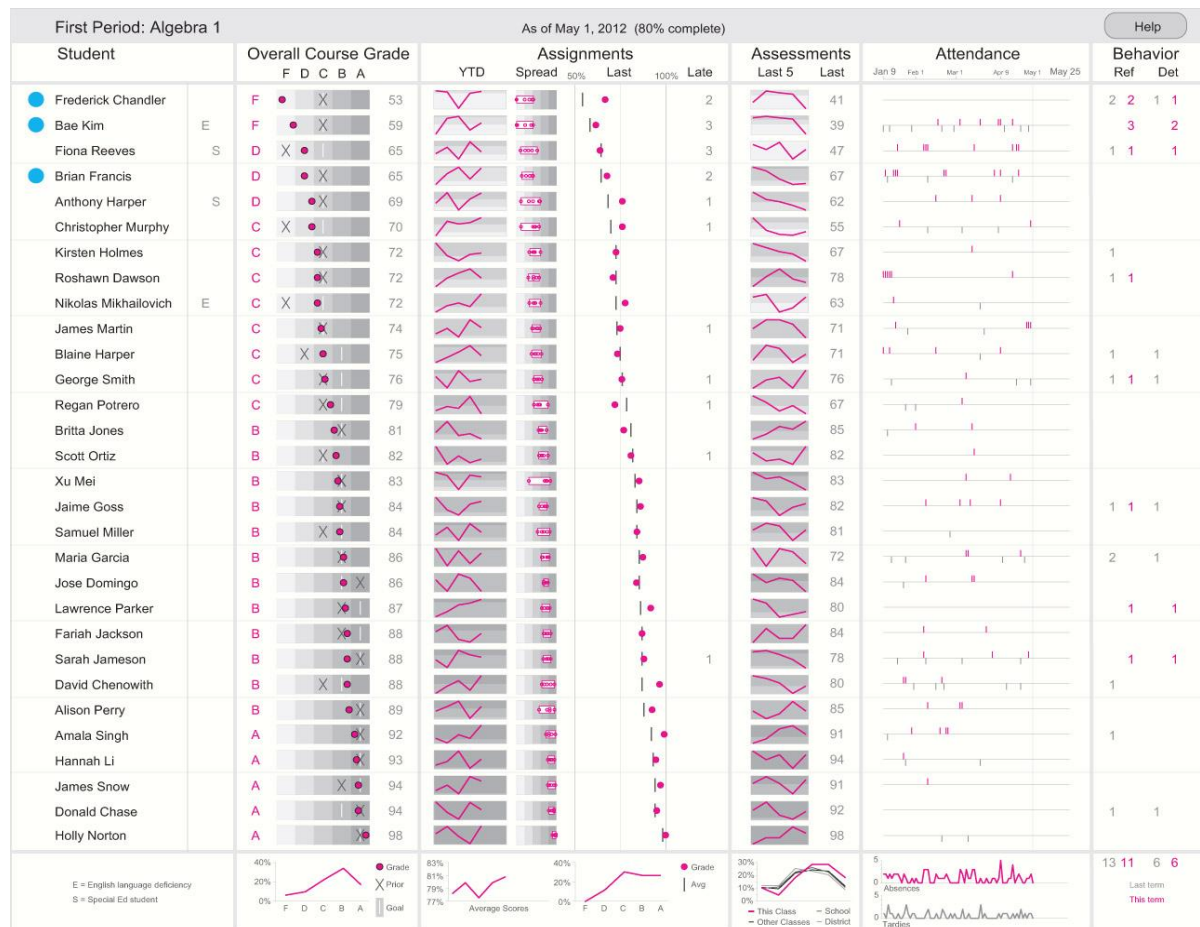


Figure 32. Visualisation of students' performance in a school class
(Source: <http://www.perceptualedge.com/blog/?p=1466>)

This dashboard represents Stephen Few's visualisation of students' performance in a school class.²² It conforms to the single-screen idea and, more importantly, provides the appropriate information to the audience—the teacher—and includes context for all measures.

What stands out is the wealth of information displayed without being cluttered; at the same time the dashboard highlights the three students who need most attention. We can also see how it integrates student- and class-level information.

Summary

The ODM dashboard may especially take notice of this for any visualisations that include all EU member states.

²² Stephen Few (2013). Information Dashboard Design: Displaying data for at-a-glance monitoring, Second Edition, Analytics Press.

5.2.2 The Open Data Institute's Company Dashboard

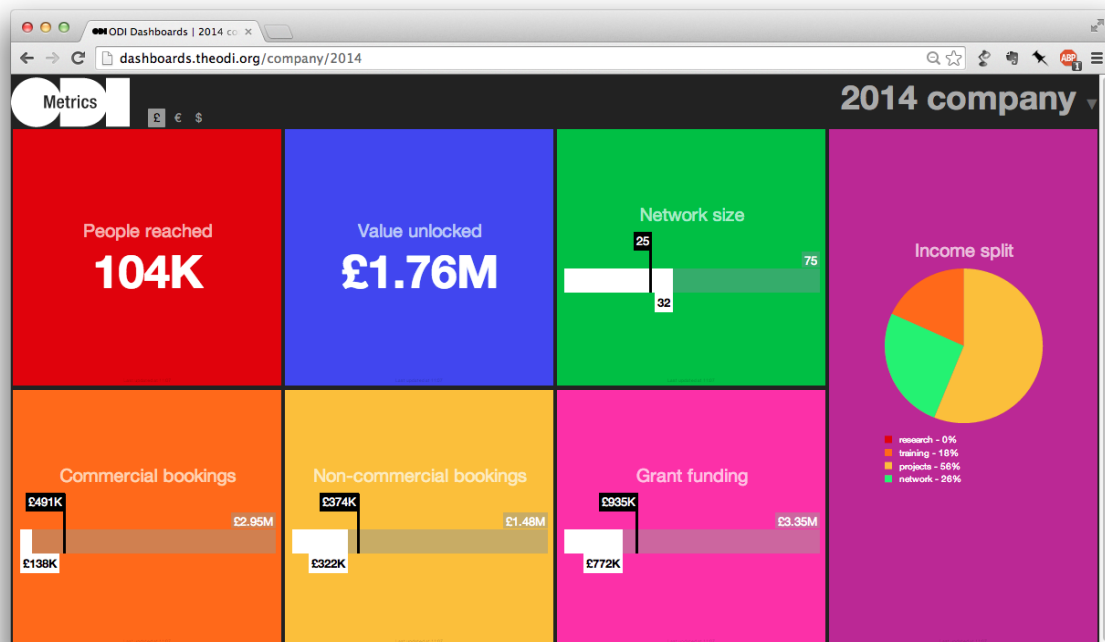


Figure 33. Open Data Institute Company Dashboard
(Source: <http://dashboards.theodi.org/company/2014>)

The ODI's dashboards exist for at least two reasons:

1. They provide a constant *state awareness*, because the dashboards are displayed online and in the office.
2. They are used in presentations and reports as a reference for ODI's work and progress.

One of the core visualisations is a variation of the bullet graph (Few, 2013) in tile 3, 5, 6, and 7. The white bar represent the current progress. The black line is the year-to-date target and the grey bar is the end-of-year target. The online version displays further information of what the metrics mean. Ideally, this contextual information is visualised directly.

The dashboards are kept simple and follow the ODI brand guidelines. Other versions of the dashboard may include historical data for context, customised ODI nodes, e.g. regional dashboards, or variations that distinguish between internal and external use.

Summary

The ODM dashboard may adopt a simple design and include additional information when mousing over the charts.

5.2.3 The Open Data Index

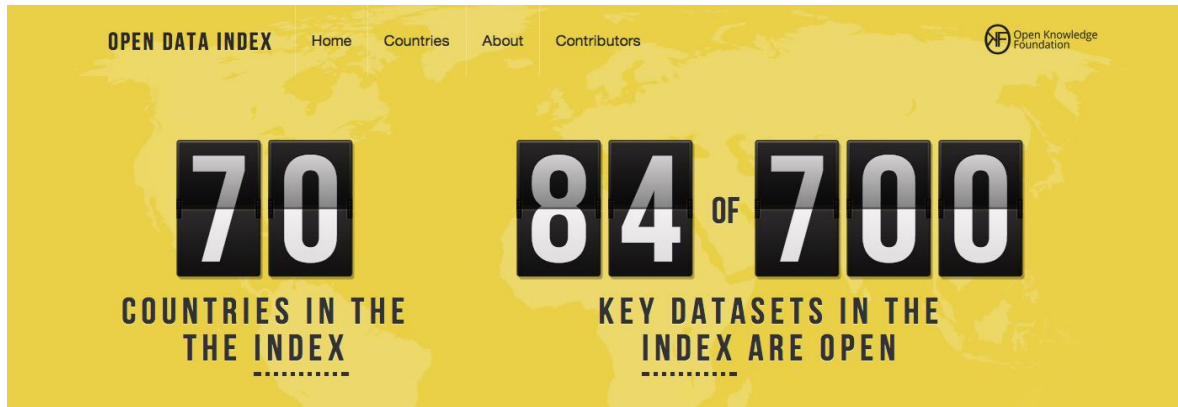


Figure 34. (Source: <https://index.okfn.org/>)

The Open Data Index by the Open Knowledge Foundation is a minimalist dashboard. It shows the number of countries surveyed and the number of dataset that are open. The latter metric provides context (“of 700”), whereas the number of countries is omitted (perhaps because we can assume it to be known). The dashboard works without a visualisation as the designers have chosen not to show the trend over time. Thus, it does one thing only: inform visitors about the index’ overall state in one glance.

Summary

The ODM dashboard may display a number on its own.

5.2.4 The London City Dashboard



Figure 35. London City Dashboard
(Source: <http://citydashboard.org/london/>)

The London City Dashboard is an example that can be improved in several areas. While it shows an impressive amount of different open data sources in one dashboard, the overall design and intention remain unclear.

For example, the design is not responsive; what you cannot see is that several Twitter feeds are truncated or not shown because they are at the bottom. It therefore violates the principle of a single screen.

Context is one of the critical considerations for dashboard design. If the audience are decision makers such as staff from the Greater London Authority, arguably the primary screen real estate, the first row, may be key indicators and not weather. Tourists may be more interested in weather, but may not care about many of the more London-specific metrics. A specific example: what is the context for the river level, should we get worried at 5.19 metres? Moreover, the colour blue does not imply an immediate signal such as red and green do.

The real-time update counter in seconds (top right for each widget) introduces distraction by blinking countdowns. We can also find an instance of excessive detail: do we really need the FTSE 100 Index at six significant digits?

We like the conservative use of data visualisations, mostly in the form of numbers and heatmaps. The dashboard also provides some interactivity through links.

Summary

The ODM dashboard may be parsimonious in its use of colours, charts and metrics.

5.2.5 Eurostat Regional Statistics

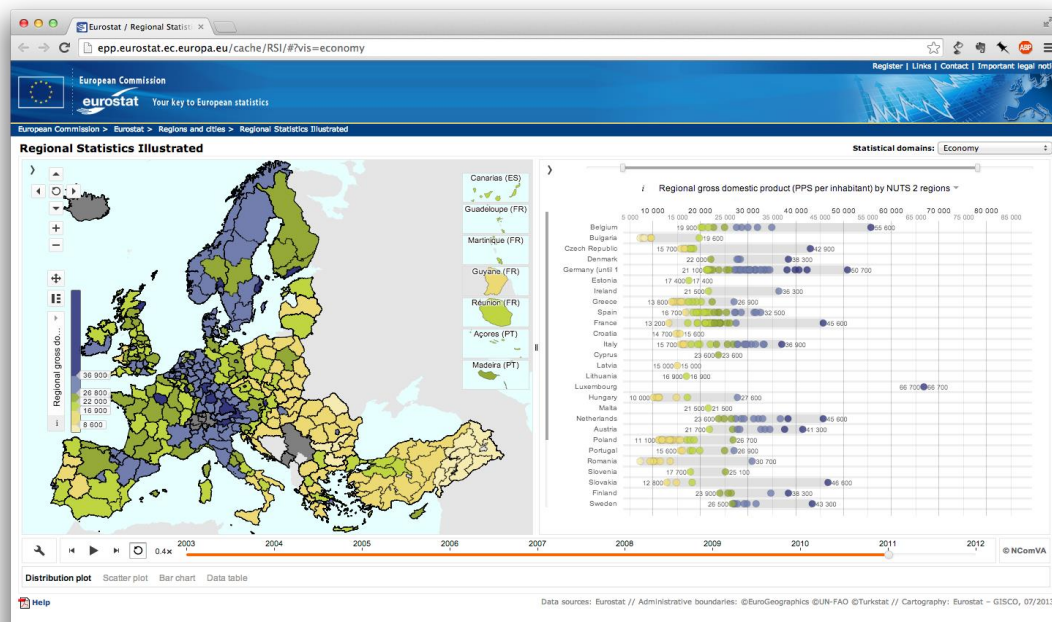


Figure 36. Eurostat regional statistics explorer
(Source: <http://epp.eurostat.ec.europa.eu/cache/RSI/#?vis=economy>)

Eurostat's visualisation of regional statistics is a remarkable tool. It fits a vast amount of information onto one screen, while preserving an overview and the ability to make meaningful comparisons. The use of colour is subtle, the integration of changes over time solved by an interactive timeline (at the bottom) and it includes many powerful filters and options for the user to explore the data further. It is an outstanding tool for data exploration – however, it fails as a dashboard because the main requirement *information at a glance* is not in its primary scope.

Summary

The ODM dashboard may remain a dashboard and not become a data exploration platform.

5.3 Dashboarding requirements of OpenDataMonitor

5.3.1 List of dashboards

Given the breakdown of the metrics in the section on *Metrics and Key Figures*, one set of dashboards mirrors those **levels of aggregation**. They are:

1. the high-level dashboard (aggregate, overall view)
2. per-geography dashboards
3. per-catalogue dashboards
4. per-dataset dashboards

A second set of dashboards brings context by means of **comparison**. Examples of comparative dashboards are:

1. dashboards for comparing the 28 EU member states
2. dashboards for comparing specific metrics such as open data formats
3. dashboards for comparing different data views, e.g. metrics split by type of data catalogue software platform.

5.3.2 Contextual information

The information visualised in a dashboard may require some context, possibly derived from metrics on a higher level of aggregation. For example, the dataset size in a per-dataset dashboard view may benefit from a comparison to the average size in the catalogue.

6 References

- Bertin, Jacques (1967) *Sémiologie Graphique: Les diagrammes, les réseaux, les cartes*. Gauthier-Villars. Paris.
- Cleveland, William S. (1985) *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey, USA, 1985
- Cleveland, W.S. and McGill, R. (1984) Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79:531–554, 1984.
- Few, S. (2004). *Show me the numbers: Designing tables and graphs to enlighten* (Vol. 1, No. 1). Oakland, CA: Analytics Press.
- Few, Stephen (2013). *Information Dashboard Design: Displaying data for at-a-glance monitoring*, Second Edition, Analytics Press.
- Neurath, Otto (1936). *International Picture Language*, London: Kegan Paul, Trench, Trubner & Co.
- Robbins, N. B. (2005). *Creating more effective graphs*. Hoboken, NJ: Wiley-Interscience.
- Tufte, Edward R. (1983) *The Visual Display of Quantitative Information*. Graphics Press. Cheshire, CT, USA.
- Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*. doi:10.1016/j.giq.2013.10.011