OpenBudgets.eu: Fighting Corruption with Fiscal Transparency

Project Number: 645833 Start Date of Project: 01.05.2015 Duration: 30 months

Deliverable D5.2

User Requirement Reports Second Cycle

Dissemination Level	Public
Due Date of Deliverable	Month 14, 31.07.2016
Actual Submission Date	July 29, 2016
Work Package	WP5, Test Beds and Evaluation - Journalism
Task	T5.1, T5.2, T5.3
Type	Report
Approval Status	pending
Version	1.0
Number of Pages	47
Filename	Deliverable 5.2 H2020 OpenBudgets

Abstract: This document describes several gaps, problems and road blocks experienced by journalists when accessing budget data. Uses cases, workshops and a curriculum is examined to resolve those gaps.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

History

Version	Date	Reason	Revised by
1.0	Jul 13	First version	Cecile Le Guen

Author List

Organisation	Name	Contact Information
J++	Nicolas Kayser-Bril	nkb@jplusplus.org
J++	Anne-Lise Bouyer	annelise@jplusplus.org
OKFDE	Anna Alberts	anna.alberts@okfn.de
OKFDE	Eileen Wagner	eileen.wagner@okfn.de
OKFDE	Helene Hahn	helene.hahn@okfn.de

Reviewer

Organisation	Name	Contact Information
OKI	Cecile Le Guen	cecile.leguen@okfn.org

Executive Summary

The second iteration of the gap analysis confirmed most of the findings of the first iteration and let us generalize them both in space across Member States of the European Union and within newsrooms, as we shifted our focus from expert journalists to all journalists.

Open budget data, like open data in general, cannot be used for the purpose of journalism without linking it to other resources. These resources are of two types: informational, as journalists need to access other kinds of public information before they can work with budget data ; and educational, as most of the foundations for working with budgets on the one hand and data on the other hand are mostly inexistent in European newsrooms.

The other tasks of this work package, which are developed in parallel to the gap analysis, show that tutorials, trainings and tools have a profound impact on the abilities of the journalists consuming them and will eventually enable OpenBudgets.eu to reach its full potential regarding journalists reusers.

Table of contents

1 Introduction

This work package applies the concept of transparency through open budget data, which is at the core of OpenBudgets.eu, to the field of journalism. Journalism is understood as the activity of deriving information from data and other observations, creating content with it and communicating this content to an audience. While this work package addresses all sorts of journalistic activities that can be done with budget data, a specific attention is devoted to journalism as a driver of transparency.

Recent research suggests that the link between transparency and open data, all the more budget data, is tenuous, at best. A review of 26 cases of alleged corruption by TACOD, a consortium of researchers co-financed by the Prevention of and Fight against Crime programme of the European Union, found that open data played a minor role only 12% of the time (compared with over 30% for investigative journalism)[1]. Meanwhile, public information played a role in over half of the cases, sometimes a major one. This study shows that open data alone cannot have an impact on either corruption or journalistic practices. Instead, open data must be seen as part of a larger ecosystem, closely related to the right to information or right to access public documents, as Katleen Janssen wrote made clear (2012). Research from the United Kingdom made a similar point regarding public spending data, showing that publication of information, even at a detailed level, did not lead to much reuse (Worthy, 2015). A survey of reuse of open data showed that only 14% of the surveyed administrations considered their data to have an “average” or “high” level of reuse (there was no mention of “very high” reuse).

In light of this research, our work package continued to use case studies to understand the nature of the current and potential links between open budget data, public access to information and journalism. The case studies made use of all possible information sources to answer journalistic questions and highlight the role that budget data could play in this mix.

As part of our mission to find out what could be the interactions between journalists OpenBudgets.eu, we continued to explore how journalists used budget data in their daily work. The methodology that we set up in the initial stages of the project, where every task is linked to the other through a dynamic feedback loop, was kept intact. Using interviews and case studies (T5.1), we drew a list of gaps to be bridged for journalists to be able to work seamlessly with budget data. The workshops (T5.2) let us gather more feedback as well as prepare tutorials to make journalists more efficient when working with budget data. Finally, the tools (T5.3) are scalable solutions to the gaps we identified. The detailed methodology was outlined in the first deliverable and was not altered[2].

During the second cycle of gap analysis, we expanded our focus both geographically and professionally, moving to new countries and to journalists with little to no experience working with data. We also opened a third window on gaps faced by journalists, after data acquisition

[1] TACOD, Revolution Delayed.

[2] See D5.1 User Requirement Reports - First Cycle

and data interpretation, namely communication of stories based on budget data.

2 Gap identification and resolution

2.1 Overview

The 15 gaps identified in the first cycle and summarized in deliverable 5.1 continued to be found in our research, whether in interviews, case studies or workshops. In particular, the availability of data remains the largest problem to be overcome by journalists when working on budgetary issues. The case studies of month 6 to 12 took the form of journalistic questions to which work package members or workshop participants tried to find an answer to. The journalistic questions had to do with assessing the cost of certain policies, such as subsidies to professional football, the implementation of a national policy at the local level and the amount of co-financing of European funds.

Local administrations analyzed by country during workshops or through internal research

	France	Belgium	Switzerland	Germany	United Kingdom	Greece
M 1-6	10	-	6	-	-	-
M 6-12	15	1	-	1	2	6

In addition, we conducted two full-day observations in newsrooms in Germany to understand how local journalists work with budget data.

2.2 Access to budget data

G5.1 Unavailability

Despite promises of “Open Government” and “Open Data” initiatives, our research shows that accessing budget data in a timely manner remains close to impossible in the regions we surveyed, from the United Kingdom to France, Switzerland, Germany and Greece.

National legislation can be ignored at the local or national level without any second thought and with a chutzpah that can be unsettling. In France, the law specifies that reports of the municipal councils, which contain the decisions on budgets, have to be published within eight days, for instance. But a local journalist and activist says:

> *Most of the time the delay is not respected. When we finally succeed to obtain a public reports it’s too late. If we want to communicate about it, no one will be interest as it’s not a hot topic anymore. Worst, it’s too late to act and contest the points in the municipal council's resolution as the deadline for appeal (2 months) is expired.*
>
> – A journalist and activist in Bordeaux, France.

Still in France, the legislation requires local councils to publish their decisions on the internet. Participants in the Lille workshop showed that at least one such council (Communauté d'agglomération de Montbeliard) does not publish any document online. When asked for the documents, the local administration asked the participants to come down to the local administration's main office and request the documents on paper. This situation is not limited to France. In the United Kingdom, the city of Sunderland offers no document on its website.

It must be noted that not all administrations are like that. Coventry, in the United Kingdom, did provide documents, as did many other local administrations in the surveyed regions. However, access to the deliberations of the local council often do not suffice for a timely analysis of local budget data, as the other gaps we assessed show.

Another worrisome factor concerns privileged access to budget data. A journalist in Germany, who maintains a good relationship with the city administration, claims they receive all budget data directly on a CD. The documents on the CD are complete and updated, but nowhere to be found online.[3] This suggests that the data – at least in this form – is not published by default, and not accessible to all journalists.

Beyond accessibility, a final problem was identified during one of the workshops (Greece). There, participants were reluctant to ask the administration for data because they expected the administration to refuse to communicate documents. Such a situation creates a negative feedback loop as journalists integrate the unwillingness of the administration to communicate budget data and therefore shy away from reporting on issues involving budget data, therefore making it easier for corruption to settle in and prosper undiscovered.

G5.2 Refusal to publish

Building on the first gap, administrations often go beyond not publishing budget data and actively refuse to communicate it, once again in a totally illegal manner. During a workshop in Lille, France, a participant who was looking for the list of subsidies given by a public administration was told by the Région Bretagne that "subsidies are only a matter between the region and the beneficiary. It is not a public matter". Of course, subsidies are a public matter ; the list of beneficiaries is voted by the regional council and should be, by law, available on the website of the institution.

An expert journalist from Spain also pointed out that local administrations refused to communicate documents. However, he did this research at a time when the local freedom of access to public documents legislation had not yet been passed. He expected that the new legislation would ease his relations with local administrations when requesting documents.

As was already pointed out in the first report, employees of local administrations have no idea what the legislation on public access to documents states. In a workshop in Paris, a civil servant at the city of Saint-Ouen said to a participant "I don't know if I have the right to communicate the

[3] Interview with journalist from Donaukurier, Mar 30th 2016.

[budgetary] documents you're asking for". Of course, the civil servant had not only the right, but the duty to do so.

Finally, public administrations often take cover under a possible "commercial secrecy" clause when refusing to communicate documents.

> *The city of Bordeaux refused to communicate the contract of the new stadium's construction and chose to communicate on the construction costs only, keeping the operational costs over 30 years hidden. We had to bring personal contacts to bear to obtain the financial model and finally be able to do the provisional budget.*
>
> – A journalist in Bordeaux

G5.3 Format

No new aspects of this gap were uncovered during months 6 to 12. All previously noted research remains true: Almost no public administration provided comprehensive budget data in a machine-readable format (only the city of Bonn, Germany, did). Structured PDFs were the best format workshop participants and researchers could hope for, unstructured PDFs were the rule.

G5.4 Timeliness

As noted by one of the Bordeaux journalists above, accessing documents in good time can be hard. Another aspect of this gap that was uncovered during recent research regards the archives of public budget data. Administrative reorganization, which happens fairly regularly among member states, can lead to loss of data. It is the case in France, where the "régions" were merged in 2015. Data from some former regions has been removed from the internet as websites disappeared. Région Franche-Comté, for instance, disabled its website after reorganization of regions in France, as a participant in the Lille workshop reported.

This difficulty in accessing archives in a usable way was confirmed by an expert journalist from Switzerland, who said that it was very hard to get comparable data from beyond the previous year. "Administrations always compare data year-on-year, but always tell you that data cannot be compared with previous years, because of methodology issues," she said.

G5.5 Completeness and level of details

The points raised in the first cycle report hold true: Budgetary documents that are released are of little interest to journalists. They hold too little relevant information that can be reused effectively in a written article or in a audio or video report. Most of what is actually used by journalists comes from press releases or phone interviews with administrations and publishing preliminary buget data online does not change this.

Sometimes data is presented in a pre-analysed fashion, meaning that in the absence of details certain narratives are suggested to journalists without giving them the full, unbiased picture.

(This was the case with a “spending review booklet” that was distributed to journalists in Ingolstadt, Germany.) What can be used in stories are detailed data, which includes the beneficiary of a spending item or the precise source of a revenue. Such information can only be collected from manual research through the city council documents. An expert journalist from Switzerland we interviewed explained that she had to go on location and raise voice to get paper documents showing the revenues from a specific publicly-owned company. (She did an investigation in the dependency of local administrations on hydroelectric plants, which can represent up to 30% of a city’s budget).

Some exceptions should be noted: There are public administrations (such as the Région Provence Alpes Côte d’Azur in France) that pro-actively publish lists of subsidies down to the beneficiary level. The beneficiaries of European funds such as the Common agricultural policy or the Structural funds are also easily identifiable thanks to the national or regional portals that were built to this effect.

Importantly, it must be noted that detailed preliminary budget data are of little use. The accounting documents used by cities, for instance, can be very detailed (a city in Germany publishes the full 170,000 lines of its budget in Excel format). However, without analytical accounting (see below) or access to the spending data, such detailed documents are barely usable by journalists.

G5.6 Identification of private beneficiaries

As was noted in the first cycle of the gap analysis, public administrations routinely mix beneficiaries of public spending, lumping together the entities of a larger whole. A participant in the Lille workshop noted that is was especially hard to identify which of the non-profit and for-profit entity of a football club received a subsidy.

G5.7 Identification of co-financing

We pointed out in the first cycle of the gap analysis that it was impossible to precisely identify money flows through several administrations. OpenBudgets.eu partners UEP proved the point once more by looking for the European structural funds spent in the city of Debrecin, Czechia[4].

G5.8 Lack of standardization

The gap that we identified in the first cycle of the analysis was underappreciated. We showed that different public bodies had different accounting standards, which made it difficult to compare their budget data. However, this was an understatement. The differences in accounting, especially between cash-based and accrual accounting, can make any comparison not just complicated, but plain wrong. A cash-based accounting system, for instance, does not take into account provisions for asset depreciation (e.g upon the completion of the building of an asset, an amount representing a share of the building value of the asset is added to the

[4] See their blog post *Tracing EU funds, a case study*.

expenditures of the income statement to mark the future need for renovations) or for future liabilities (e.g pensions). The difference between the two can mean that a public administration might show a positive net result of several million euros using cash-based accounting and a much larger deficit using accrual accounting. Thus, comparing performance between administrations without taking into account their accounting methods is more than meaningless: it is misleading. This difference was explained in an article published on the OpenBudgets.eu website.[5]

Even when a public body makes the switch towards accrual accounting, as most have, an absence of unified guiding rules can allow some institutions to create their own accounting system. The University of Heidelberg, for instance, shaped its financial reporting system according to its own preferences. The pension provisions of the public official and the fixed assets are not included in its financial statements. As a result, even financial documents created under an accrual system can be very impossible to interpret without a solid knowledge of their underlying accounting practices.

The lack of standardization can also be an issue *within the same public body*! A workshop participant in Strasbourg found that in Saint-Etienne, France, the city booked the exact same expense, a communication contract given to the local football club (so that the city's name appears on their jerseys) under an investment expenditure under one legislature. The following legislature booked it under operating expenditures.

Finally, the hierarchy of territories can be non-standard within the same country. In the United Kingdom, territories are divided in regions, counties, districts and boroughs. However, the order of the subdivisions is not coherent across the territory of the country, making it difficult for researchers to know what public institutions operate over a given area.

G5.9 Contradicting data sources

The case studies continued to yield examples where official data sources contradicted each other. In Coventry, United Kingdom, for instance, an ERDF subsidy went from £4.7m to £4.3m depending on the source, without any explanation given by the local authorities. Absent access to the actual spending data of all parties involved, finding the answer to this question was impossible (the European Commission and the Coventry City Council do disclose detailed spending data, but the other parts to the project being built with the subsidy did not).

2.3 Understanding budget data

G5.10 Understanding basic terms and context

Observations in German newsrooms have demonstrated the need for proper support and tools for budget and spending reporting. Because budget discussions only appear two to three times a year (regionally and nationally), and because resources are generally sparse for long-term

[5] See Public sector accounting in Europe

investigative work, it is uncommon to have designated budget and spending reporters in house. This is why reporting happens sporadically and without much direct contact with budget or spending data. Journalists have repeatedly stressed the need for standardised, on-time, and pre-analysed data that would fit into their fast-paced daily research routines.[6]

Along similar lines, a participant to a workshop in Strasbourg said: "I don't trust myself to understand budget data". Budget data is intimidating and many journalists do not trust themselves to read source documents directly. However, the tutorials that we wrote as part of T5.2 have proven effective. Between the first workshops (in Neuchâtel and Strasbourg) where we did not offer a paper version of the tutorials, doing instead a brief presentation, and the three that followed (Lille, Paris, and Thessaloniki), we noticed a marked improvement in the way participants worked with budget data. The two-hour crash course in public-sector accounting works and will continue to be refined in the coming months. In Lille, a participant even said that “[she] had no idea public documents could yield so much information” and that she would use them more in the future.

An unsolvable issue was however raised during the workshop we organized in Thessaloniki. There, participants did not consider public spending as a category in itself. Instead, they considered the activities of public bodies within the wider frame of the activities of the people who controlled them and made no difference between a private company and the public bodies operated by the owner of a said private company. While this may be seen as a misconception to be corrected, it is not. Several oligarchs actually control municipalities in Greece and do not hesitate to fund what other countries consider as public works (the development of a public park, for instance) with their own capital if the city council cannot afford it. The city of Piraeus is a case in point. There, local oligarch Vangelis Marinakis was on the ballot in 2014 for the city council and made no secret of his intention to run the city behind the scenes. Once elected, he stated that he financed construction of squares, parks, children’s playgrounds and sports facilities, at his own cost. The very concept of a public budget must therefore be understood differently depending on the level of interconnection between personal and public activities.

Perception of public bodies activities and how public money is spent may vary across local, regional or national political contexts, but in general the understanding of budget data remains mainly unappreciated.

G5.11 In-kind spending and gifts

Non-financial transfers from public authorities to private beneficiaries continued to appear in our research. It represents the hardest gap to bridge regarding the data structure of OpenBudgets.eu and OpenSpending.

The city of Bordeaux, for instance, decided not to charge the private company involved in the new stadium’s construction for the local taxes.

[6] Observations and interviews conducted in two selected German newsrooms, Mar/Apr 2016.

> *In the provisional budget we did for the stadium's construction, we counted €79.6m in tax exemptions as an in-kind gift from the city. It makes sense since it is the same operation as if the private company paid the taxes and then the city gave it back. It is a real money flow to take into account.*
>
> – Local journalist, Bordeaux

However, proponents of the stadium, in this example, will point out that the stadium would not have been built at all, were it not for the €79.6m tax rebate. Other examples point to the complexity of the task. A specific tax in France is levied over public entertainment. However, every city we analyzed waived it for football matches. Therefore, it is hard to consider the non-payment of such a tax as an in-kind gift, as every agent part to the exemption assumes that the tax will not be paid due to habitual practices. In the same vein, the price of land is extremely hard to assess, because, more often than not, there is no market for the transactions that journalists need to analyze (the tract of land used for a stadium, for instance, is often much bigger than any transactions that took place in the area).

To precisely estimate the value of in-kind spending, a robust framework would need to be created, which would offer guidance to those doing the estimate. However, creating such a framework would be beyond the goal of this work package, not least because journalists would not use it (see below our points on workshops and communication).

G5.12 Off-balance sheet operations

Our research continued to yield examples where off-balance sheet operations, that is to say liabilities that the public administration takes without it showing in its accounting, play a major role. In Spain, for instance, an interviewee stated that public subsidies can take the form of a special schedule for the repayment of tax and social contributions arrears. Such special arrangement can be made of hundreds of millions of euros, but not be visible in public budget data.

The importance of off-balance sheets operations reinforces the importance of gap 5.1 (access to documents), as the details of such deals are to be found in the archives of the local councils. Furthermore, it reinforces the need outlined above for a clear framework for the valorisation of non financial transactions. A city can offer to act as a guarantor to a private entity contracting a loan, for instance (as participants in the Lille workshop found out). In such cases, the value of the off-balance sheet engagement is hard to assess.

G5.13 Understanding of accounting legislation

Even when the basics of public sector accounting are understood, some of the material journalists face during an investigation can require expert knowledge. An interviewee from Switzerland explained that it can be hard to understand difference between budgets and spending, for instance. She continued: "Experts say that something is 'normal', but it's hard for a journalist to understand if this such feeling of normalcy is justified. The Swiss Confederation, for instance, posted a positive result of 2 billion Swiss Francs, but kept saying it cannot finance social programs”. The solution this journalist found is to ask university professors for advice.

Other examples were found during the research. Public-private partnerships especially usually contain specific clauses that can vary a lot and imply hidden costs for the public administration. To assess these costs, one must understand the above-mentioned clauses in details. The hidden costs can only be calculated and evaluated if we understand:

- The extent of the responsibility of the public party to the partnership. Some contracts have an assignment agreement in respect of rents receivable which means that the public administration is responsible of the loans contracted by the private partner.
- The interest rates that must be paid by the public administration on the loans taken for the construction of the asset by the private party, to which the public party is the guarantor (e.g variable rates).
- The compensation fees that have to be paid by the public administration and under what conditions. In Bordeaux, for instance, the public-private partnership mentions that the city has to pay compensation fees to the private company in case of the contract is declared null or cancelled by a judge.

Most of this technical knowledge will be communicated to journalists through the tutorials that will be written and published online.

G5.14 Analytical accounting

The lack of analytical (or mission-based) accounting makes any investigation into a specific aspect of a local community extremely time-consuming. The case study on public subsidies to professional football showed, for instance, that based on currently available budget data it would be extremely costly to assess the real amount spent on the maintenance of a stadium by a public body. One could, if it were available, list all the purchases done over a year using public procurements (for instance new paintings, purchase of synthetical grass, purchase of gas for heating the stadium etc) but such purchases are rarely associated to a specific entity like a stadium.

In Spain, an interviewed journalist highlighted how this gap was linked to gap 5.2. There, the government does do analytical accounting on subsidies to football clubs but refuses to disclose the breakdown by club, arguing that such matters are protected by privacy law.

A case study in Bonn, Germany showed that analytical accounting by the public entity could dramatically speed up the work of a journalist. The journalistic question we set out to answer was the expected cost of a specific measure that the city was undertaking. The detailed budget data provided by the city did not help in answering the question, because the official ontology used by the city’s accountants assigned different types of spending to different accounts, even if the spending items were related to the same mission (no analytical accounting). However, the city provided us with one document of analytical accounting that they themselves made, based on the features of the accounting software. This document enabled the journalists taking part in the case study to quickly check the assumption made in the budget and to answer the journalistic question (the total cost of the measure was estimated by the journalists of the case study to be between €3 and €19 million). Interestingly, the city itself published its forecast of the cost of the measure after the case study was completed. It estimated its cost to be €16 million, right between our estimates. This case study showed that good budget literacy, obtained through tutorials, coupled with access to some budget documents of an administration, can enable journalists to answer journalistic questions efficiently.

G5.15 Difference in definitions

No further example of this gap have been found, it can therefore be removed from the list.

2.4 Working with budget data

Among journalists is a misunderstanding not just what budgets are, but also what datajournalism entails. datajournalism, especially with the context of investigative journalism, is often misunderstood. Aspiring data-journalists who we have spoken to at workshops for the Football Tax and conference workshops on datajournalism skills have a strong focus on the tools and underestimate the amount of manual and traditional journalistic work that needs to be done to eventually turn data into a story with strong visuals presenting the result of any analysis. In our previous analysis of financial-datajournalism workshops and learning materials available, we found that the combination of financial-journalism and datajournalism training should be developed in this project.

We have continued to work on the observation of datajournalism workshops and training, and continued to work on the gap of financial- and datajournalism training. New insights have shown us that teaching tools does not suffice. Datajournalism and its tools are part of larger journalistic work processes. Although tools and automatization may sound like they will elevate work, in the short run it means more work, especially when used outside of the context of investigative journalism but for normal day-to-day reporting on politics. As part of the larger investigative journalistic workflow, it is a mere necessity to work with structures and certain tools can indeed elevate work. However, we have to remain aware that one must pick the tool appropriate within the workflow and not centralize the tools themselves.

G5.15 Integrating data journalism and financial reporting

The observations found that due to the sporadic nature of budget debates, reporting is mostly based around events. Interviews and observations in German newsrooms showed that there is virtually no room for long-term, investigative journalism around particular issues.[7]

The possibilities that data journalism offers in budget reporting have not been sufficiently explored. A major German newspaper, with an in-house team of data journalists, was surprised to be asked about assigning data journalists budget- or spending-related tasks. It is unclear whether it was the data journalism team which was not fully integrated in the work routine, or the

[7] Interview with Tagespiegel (April 13th 2016) and Donaukurier (Mar 3rd 2016).

lack of attention paid to budget and spending data that is to blame for this missed opportunity.

G5.16 The Need for Basic Tools

Some observations of journalists' workflow were highly surprising. Where we expected journalists to use at least basic spreadsheet tools, we found during one of our observations that all the work was still done by hand. A journalist in Germany, in charge of budget reporting for a medium-sized regional newspaper, generally writes down numbers from a spreadsheet and calculates sums and percentages using a portable calculator. When asked whether they had considered using a computer, they replied: "Nothing has changed in my investigation in the last thirty years; everything works well for me." A similar behavior was observed during workshops with journalism students in France.

In a discussion on the creation of learning material for data journalists the former project manager of the founding team of Open Spending and of School of Data noted that the most important tools to teach are spreadsheet operations and basic tools.

Most people will think about data journalism imagining the end-product wrapped in beautiful maps, stats, or visualisations. But most of the work in data journalism is done in spreadsheet programs, and most journalist are greatly helped by teaching basic tools on how to work with spreadsheets, pivot tables, and getting and cleaning data.

For this purpose we have developed the first guide for teaching the basics of data journalism, based on the work of Data Trainer Marco Pires and Anne-Lise Bouyer. This guide is now tested out and improved in several iterations to produce a train the trainers guide for NGO and Journalist training for OpenBudgets.eu. In following iteration the material on financial journalism will be added.

2.6 Communication of stories based on budget data

For journalists, accessing budget data, understanding it and producing a story with it are just the first steps of a longer process. The story needs to be conveyed to the audience in a way that attracts the readers' attention. It is common knowledge in the newsroom that stories on budget data, however outrageous the findings might be, do not fare well. Newsroom "common sense" argues that people do not respond to amounts that they cannot grasp. However, no research ever tried to check that assertion.

Our research on how to best communicate budget data helps us design the tools of T5.3.

2.6.1 The problem

Reports on stories involving large amounts of money rarely gain traction because of the numbers alone. The Russian Laundromat[8], for instance, exposed how $20 billion were laundered with the help from European banks. In France, a public servant was found to have filed €40,000 worth of taxi expenses[9] (500,000 times less than the Russian story). The first story received little attention, while the second was headline news. How can stories involving huge amounts be so easily ignored by journalists and their audience?

It has been shown in many experimental set ups that animals and humans alike tend to overestimate quantity from numerosity. A piece of food split in pieces is considered as more filling than the same amount of food in one piece, for instance. This cognitive bias is known as the "numerosity heuristic" (Pelham et al., 1994). The numerosity heuristic has been shown to exist in monetary amounts as well (Soman et al., 2002).
The preference of the audience for stories involving smaller amounts go against the numerosity heuristic. We offer the following explanation:

- Most people have a poor understanding of large orders of magnitude. We can easily make sense of amounts we live with (up to a life's worth of earnings) but are unable to process larger amounts.
- Because of this impossibility to make sense of large amounts, we are unable to assess the importance of a story based on the amounts involved. Stories involving small amounts, where the discrepancy between the expected amounts and the reported amount is great (the taxi story), are more outrageous than stories involving large amounts (the Russian one) where there is no reference as to what a normal amount would be.

2.6.2 Experimental setup

Users are given an interactive questionnaire, presented as a game called "Order of Magnitude Guesser". This experimental setup is inspired by GeoGuessr and EarthQuiz (De Paor,

[8] Online at https://www.reportingproject.net/therussianlaundromat/

[9] Read online *France archives boss Saal resigns over €40,000 taxi bill* http://www.bbc.com/news/world-europe-32510604

Whitmeyer, Bentley, and Dordevic 2014), which were used to assess place location knowledge (Zhu et al., 2015).

The goal is to correctly guess the order of magnitude of 10 questions. Questions ask orders of magnitude of daily life items and larger ones and provide three choices, the wrong ones being at least 2 orders of magnitude above or below the right answer. The answers can be phrased in three ways: Numerical (“1,000,000€”), written as text (“1 million”) or relative to a daily-life item (“the lifetime earning of an average employee”).

At the end of the 10 questions, personal questions are asked about age, education and occupation.

2.6.3 Hypotheses

1. The smaller the amounts, the better the guesses of participants. Daily-life items are better interpreted because participants cannot put an amount on an unknown item.
2. When answers are presented numerically, guesses are worst than when answers are written down. We assume that people cannot count the number of zeroes, even if a thousand separator delimits groups of 3.
3. When answers are presented relative to daily life items (e.g “item X costs the same as a VW Golf”), guesses are better than with the other two (amounts as number and amount as text), until the multiplying factor increases above 10 (e.g “1 billion” will be yield better results than “The lifetime salaries of one thousand average employees”). This should be true if hypothesis 1 is true. If what is needed to produce sense is a link to one’s life, it should be easy to move from one’s own life to other items of one’s life (e.g a family, a small team) but not further (a large company).
4. The rate of correct guesses does not vary according to units. If hypothesis 1 is true, the issue should not be domain-specific.
5. People with an occupation in engineering or computer science score better than others. The use of order of magnitude in daily life should help identify them in the questionnaire.

2.6.4 Expected outcome

We expect a few thousand users to play at least one game of “Order of Magnitude Guesser”. If needed, we will acquire paid traffic from Facebook, targeting the categories of users for which we lack data.

Each hypothesis will be tested using the collected data. If hypothesis 1 is true (and is backed up by hypotheses 3 and 4 also being true), we will know what is the threshold from which users cannot make sense of amounts. It will be of tremendous help for journalists communicating stories involving budget data and organizations developing tools for journalists, for they will know what range of amounts can be understood by an audience. Similarly, the outcome of hypothesis 2 will help us design better interfaces, for we will know when to use numerical values and when to use values written as text. Finally, the analysis of the data by occupation will let us know if journalists display specific behavior compared to other groups.

2.6.5 Embodiment

A final aspect of communicating budget data, which is not addressed in this experiment, deals with embodiment. Many stories involving outrageous sums of mismanaged funds tend to focus on trivial details. The Elf-Affair, in the second half of the 1990's, showed how political parties from both the left and the right had used shady deals in Gabon and other dictatorships to fund their slush funds, to the tune of billions of French francs. Despite the enormity of the claim and the fact that top public servants (ministers and above) were implicated, much of the attention focused on a pair of shoes worth 11,000 Frs (2,000€ at the 1999 exchange rate). In Italy, the affair linking former Prime Minister Silvio Berlusconi to sex parties received much more coverage than his conviction for tax evasion. A similar pattern was mentioned by an interviewee in Switzerland. She published a series of stories on a local entrepreneur who dodged 13m Swiss Francs in a tax evasion scheme, but the stories only gained traction when a minor side-story mentioning how he also tampered the wine he produced was published.

"Common sense" in the newsroom, there again, argues that stories that are most divergent from the norm will be more successful. A story showing that a politician bought shoes worth 20 times the normal value (2,000€ instead of 100€) will be more understandable than a story about politicians embezzling €200m (for which there is no norm to be put in relation). This is in line with the seminal work by Kahneman and Tversky (1979) on perceptions of gains and loss, in which they showed that people reacted much more positively to a gain of 100 on top of 100 than to a gain of 100 on top of 1100.

In partnership with UEP, we are in the process of designing an experimental setup that would let us measure the interest of users in stories that involve mismanagement of public funds. The results of this second experiment would greatly help journalists to tailor the angles on such stories in a way that might make them interesting to their audience.

3 Conclusion

The gap analysis shows that journalists face a long list of hurdles to use budget data in their daily work. It is highly unlikely that many of them will change their current workflows to move to a data-driven approach, whereby they would use budget data as the starting point of their stories.

However, the OpenBudgets platform could become the go-to place for journalists looking for budget data, which is so direly unavailable from the local authorities. Furthermore, the tutorials and tools developed as part of T5.2 and T5.3 will increase the abilities of journalists willing to work with budget data.

Summary of the identified gaps

Gap title	Action required to bridge the gap	Progress on action

G5.1 Unavailability	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Action started with project partners.
G5.2 Refusal to publish	Increase awareness of the issue among public administration bodies (collaboration with WP6). Increase awareness among public ombudsmans (collaboration with WP6 and WP7).	Development of tutorials.
G5.3 Format	Automated PDF and Excel parsing to RDF (collaboration with WP2). Increase awareness of the issue among public administration bodies (collaboration with WP6).	Impossible to parse the PDFs.
G5.4 Timeliness	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Impossible to change the habits of administrations within this WP.
G5.5 Completeness and level of details	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Impossible to change the habits of administrations within this WP.
G5.6 Identification of private beneficiaries	Use of fiscal identification numbers by public bodies (collaboration with WP1). Tutorials for journalists on how to correctly identify beneficiaries.	Development of tutorials.
G5.7 Identification of co-financing	Increase awareness for transparency among public bodies (collaboration with WP6).	Development of tutorials.
G5.8 Lack of standardization	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Development of tutorials.
G5.9 Contradicting data sources	Tutorials for journalists and public administration officials.	Development of tutorials.
G5.10 Understanding basic terms and context	Tutorials for journalists.	Development of tutorials.
G5.11 In-kind spending and gifts	Addition of in-kind spendings to the OB ontology when possible (collaboration with WP1). Tutorials for journalists.	Ontology impossible to change. Development of tutorials.
G5.12 Off-balance sheet operations	Addition of off-balance sheet operations to the OB ontology (collaboration with WP1). Tutorials for journalists.	Ontology impossible to change. Development of tutorials.

G5.13 Understanding of accounting legislation	Tutorials for journalists.	Development of tutorials.
G5.14 Analytical accounting	Tutorials for journalists.	Development of tutorials.
G5.15 Difference in definitions	Tutorials for journalists.	Gap dropped.
G5.16 Integrating datajournalism and Financial Reporting	Tutorials for journalists.	Development of tutorials through iteration in 2016 - 2017
G 5.17 The Need for Basic Tools	Tutorials for journalists.	Development of toolbox

5 References

Bouyer, Anne-Lise and Marco Tulio Pires. ‘Free Data Journalism Tools - Slide deck’ (2016). Notes taken in Perugia, April 2016.

De Paor, D. G., S. J. Whitmeyer, C. Bentley, and M. M. Dordevic. 2014. “EarthQuiz: A Crowd-Sourced Resource for Teaching and Learning Geoscience” http://www.earthquiz.net

Heravi, Bahareh. “Tips and Tricks for Data-Driven Journalism Starters” (2016) <http://www.journalismfestival.com/programme/2016/tips-and-tricks-for-data-driven-journalism-starters> Retrieved June 2016. Notes taken in Perugia, April 2016.

Kahneman, Daniel, and Amos Tversky. "Prospect Theory: An Analysis of Decision under Risk." Econometrica 47.2 (1979): 263.

Katleen Janssen. “Open Government Data and the Right to Information: Opportunities and Obstacles.” Journal of Community Informatics Vol 8, No 2 (2012)

Pelham, B.w., T.t. Sumarta, and L. Myaskovsky. "The Easy Path From Many To Much: The Numerosity Heuristic." Cognitive Psychology 26.2 (1994): 103-33.

Soman, Dilip, Klaus Wertenbroch, and Amitava Chattopadhyay. "Currency Numerosity Effects on the Perceived Value of Transactions." SSRN Electronic Journal SSRN Journal (2002).

Wehrmeyer, Stefan.“Reproducible and transparent data journlism with modern tools” (2016). <http://www.journalismfestival.com/programme/2016/reproducible-and-transparent-datajournalism-with-modern-tools> Retrieved June 2016. Notes taken in Perugia, April 2016.

Worthy, Ben. "The Impact Of Open Data In The UK: Complex, Unpredictable, And Political." Public Administration Public Admin 93.3 (2015): 788-805.

Zhu, Liangfeng, Xin Pan, and Gongcheng Gao. "Assessing Place Location Knowledge Using a Virtual Globe." Journal of Geography 115.2 (2015): 72-80.

6 Appendix: The Toolbox

Based on observations at Perugia from the training of Marco Tulio Pires and materials written by Marco Tulio Pires and Anne-Lise Bouyer and additional materials gathered at the School of Data summerschool, the following tutorials and tool descriptions are developed: tools to find and get the data, tools to clean the data, tools to analyze the data, tools to visualize the data and tools to tell a story with data. The following displays the first iteration of the trainers guide for data journalism, largely based on the 'Free Data Journalism Tools - Slide deck' by Anne-Lise Bouyer and Marco Tulio Pires (2016).

1.1 Tools to Find and Get the Data

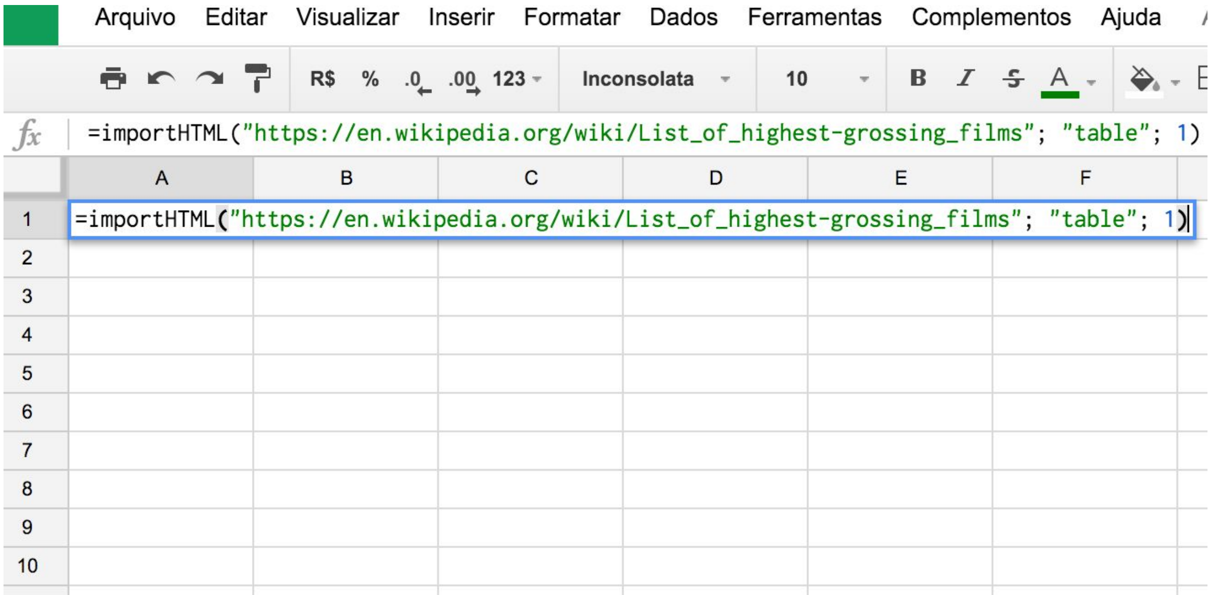
Journalistic work often starts with a question or a hypothesis relating to a particular field. Finding the right data sets is crucial when beginning a data-driven investigation. Where can one find data?

Many government institutions provide data via open data portals or repositories. Data portals as well as data sets are usually updated and added to on a regular basis by government officials. The biggest European data repository for budget data is https://openspending.org/. Freely available data sets are also provided by statistics institutes, universities, and news sources, such as http://data.worldbank.org/, https://datahub.io/, http://data.un.org/, http://stats.oecd.org/Index.aspx, and the APIs from the New York Times (application programming interfaces which can be used to request data from a large database as a source).

Under the Freedom of Information Act (FOIA), every person has a right to request any recorded information held by a public authority, such as a government department or local council. Data can also be requested via the FOIA.

Getting the data is one of the first and most important steps. The following tools can also be useful when procuring data.

Example 1: Working with Google Sheets

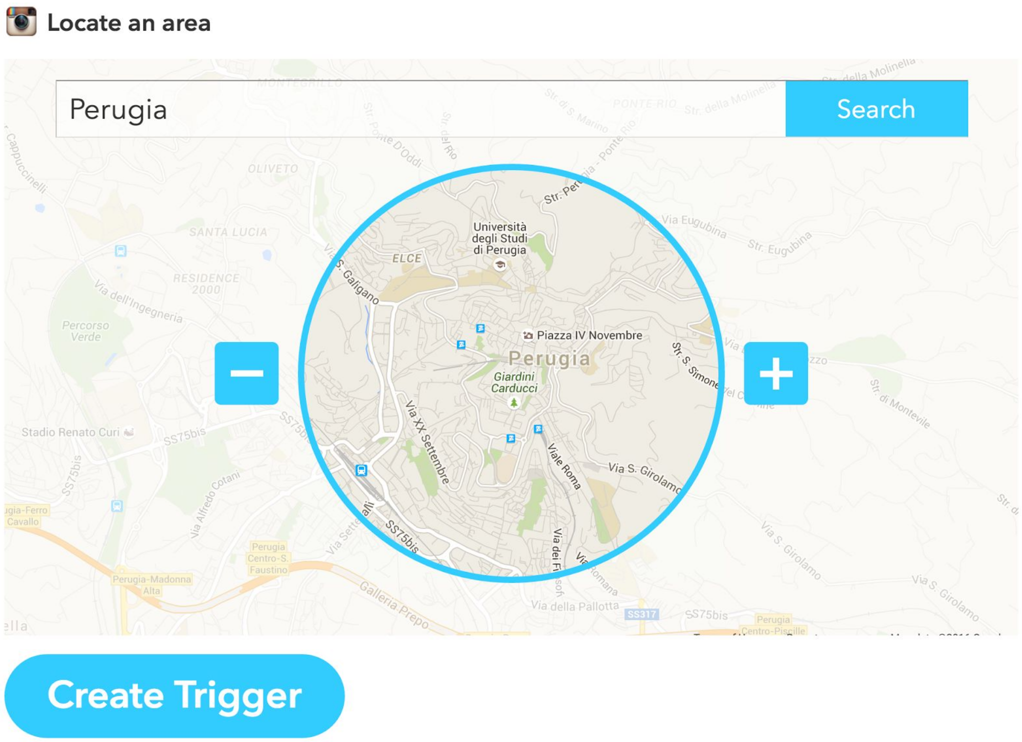


```
=importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films"; "table"; 1)
```

(Screenshot Google Sheet, Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016)

In html pages, data is structured but not in an easily usable format. One commonly used tool that can be used to get data from html pages is Google Sheets. Google Sheets is the online version of Excel that can be collaboratively used and edited. Through using the “=importHTML” function, tables and lists can be easily imported. The above example shows how to access the URL for a list of the “highest-grossing films” from a Wikipedia page. The function processes the first list in Google Sheets directly (Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

Example 3: Working with IFTTT



(FTTT Screenshot, Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016)

A useful website to collect social media information and other types of data is IFTTT (If This Then That). IFTTT connects to different services online, such as Dropbox, Instagram, Twitter, Facebook and New York Times stories. As soon as something that matches the user's IFTTT keyword search happens, it carries out a specified action (for example, collecting the data into a table). In this example, we are creating a collection of every picture posted publicly on Instagram in Perugia. IFTTT automatically saves all available information (metadata) relating to each picture in a Google Spreadsheet. In this way, journalists are able to scrape information from social media without needing to program anything. In this case, the date, username and status of the picture as well as the links and the picture itself are collected.

To sum up, if the data you're looking for is in a table on a web page, or a series of tables, where each page has a URL, use:

- =importHTML() in a Google Spreadsheet
- Web Scraper, the browser extension

In other cases, you can use:

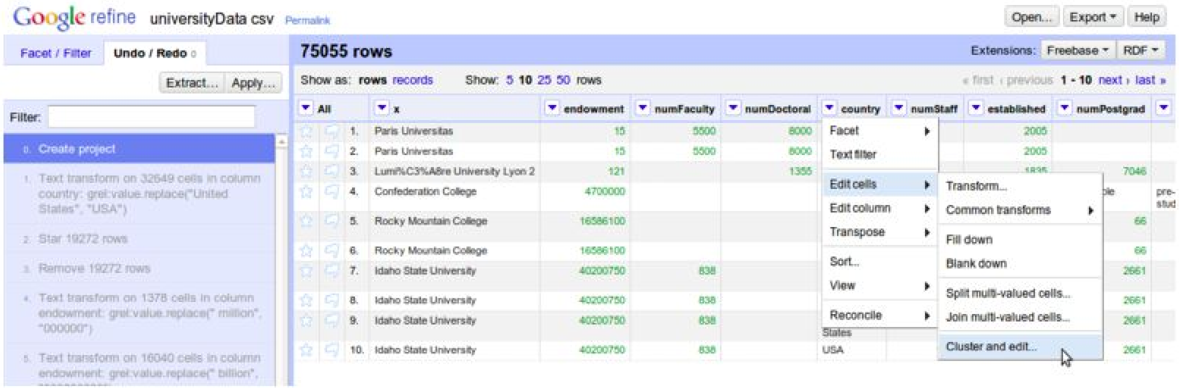
- IFTTT, to track and collect social media information online
- Or any programming language to write scrapers, like Python

(Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

1.2 Tools to Clean the Data

When data is found, it is important to clean it to ensure that data is machine readable and ready for further analysis and visualization.

Example 1: Working with Open Refine (formerly known as Google Refine)



75055 rows

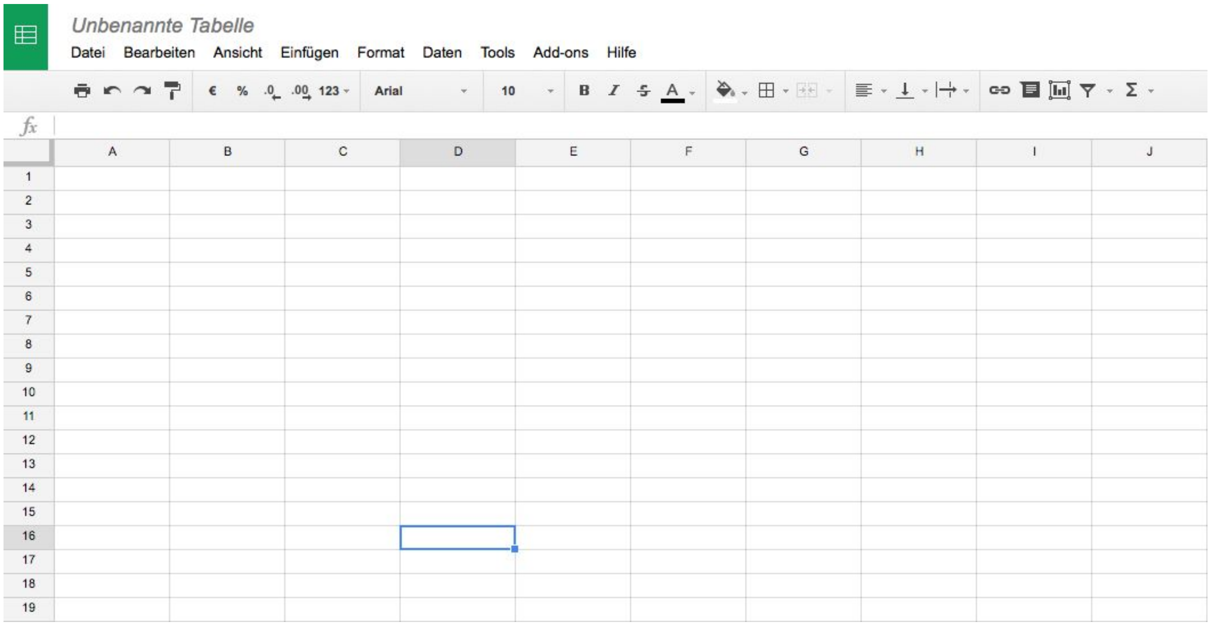
0. Create project
1. Text transform on 32649 cells in column country: grel:value.replace("United States", "USA")
2. Star 19272 rows
3. Remove 19272 rows
4. Text transform on 1378 cells in column endowment: grel:value.replace(" million", "000000")
5. Text transform on 16040 cells in column endowment: grel:value.replace(" billion",

All	x	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	
1.	Paris Universitas	15	5500	8000			2005		
2.	Paris Universitas	15	5500	8000			2005		
3.	Lumi%C3%A8re University Lyon 2	121		1355			1835	7046	
4.	Confederation College	4700000						ple	pre-stud
5.	Rocky Mountain College	16586100						66	
6.	Rocky Mountain College	16586100						66	
7.	Idaho State University	40200750	838					2661	
8.	Idaho State University	40200750	838					2661	
9.	Idaho State University	40200750	838		States			2661	
10.	Idaho State University	40200750	838		USA			2661	

(Google Refine Screenshot, Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016)

One of the best available tools for data cleaning is Open Refine, an open source software. Open Refine can be used to clean typos, reconcile and cluster different data sets. The tool offers functions for various aspects of cleaning data, such as for reconciling inconsistent spellings (e.g. ‘USA, ‘U.S.A.’ and ‘U.S.’), converting text into numeric values (e.g. $123 million to $123,000,000), extracting and cleaning data in date format and for the removal of duplicate rows. Its use is especially recommended when working with sensitive data, due to the fact that this tool can be downloaded and used offline. It is also possible to export projects from Open Refine and to share them with colleagues. (Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

Example 2: Working with Google Sheets



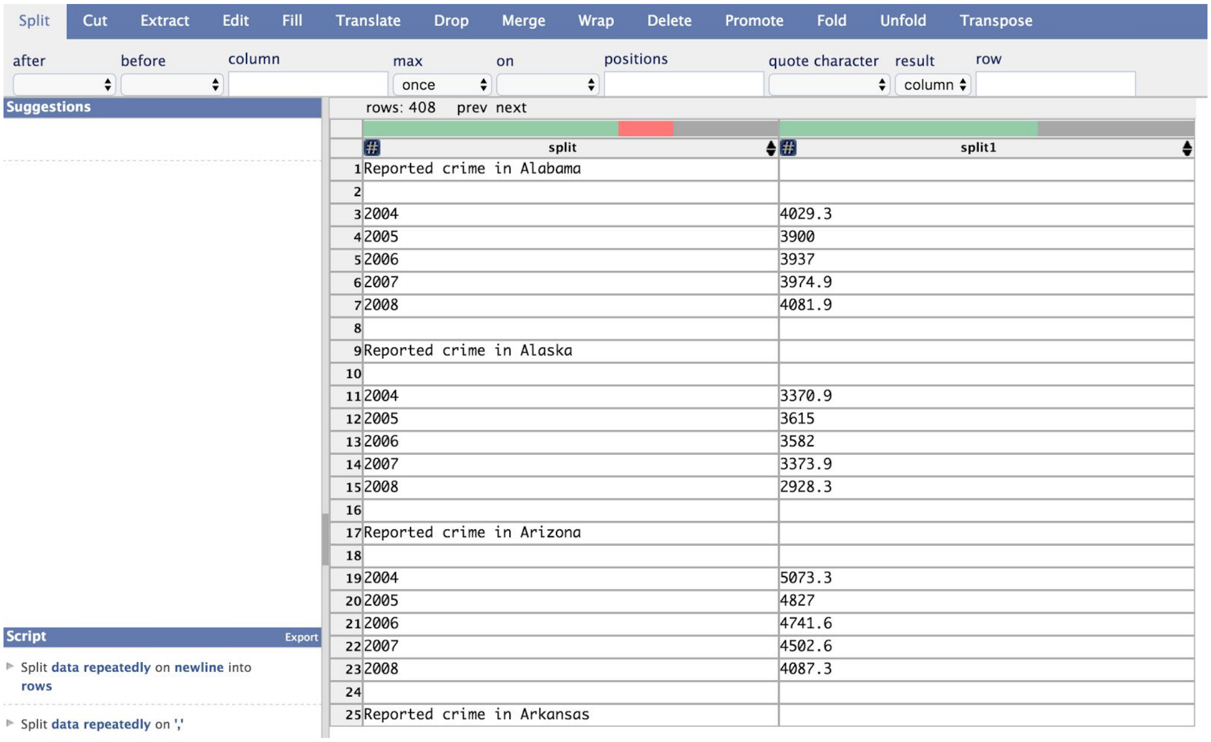
(Screenshot google sheet, Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

The formulae available in Google Sheets, or comparable tools like Excel or Libre Office, are not only useful for acquiring data, but also for cleaning data. The following formulae are especially useful for data cleaning:

- =split() -> splits text based on criteria
- =trim() -> trim leading and trailing spaces
- =proper() -> capitalizes each word in a string
- =concatenate() / =concat() -> joins strings together
- =clean() -> removes non-printable character from string
- =to_data() / =to_pure_number() / =to_dollars() / =to_text() / =to_percent() -> convert strings
- =googletranslate() / =googlefinance() -> online services like translations and conversion of currencies and many more
- find and replace function -> for example, to clear dots

(Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

Example 3: Working with Data Wrangler



after: | before: | column: | max: once | on: | positions: | quote character: | result: column | row:

rows: 408

	split	split1
1	Reported crime in Alabama	
2		
3	2004	4029.3
4	2005	3900
5	2006	3937
6	2007	3974.9
7	2008	4081.9
8		
9	Reported crime in Alaska	
10		
11	2004	3370.9
12	2005	3615
13	2006	3582
14	2007	3373.9
15	2008	2928.3
16		
17	Reported crime in Arizona	
18		
19	2004	5073.3
20	2005	4827
21	2006	4741.6
22	2007	4502.6
23	2008	4087.3
24		
25	Reported crime in Arkansas	

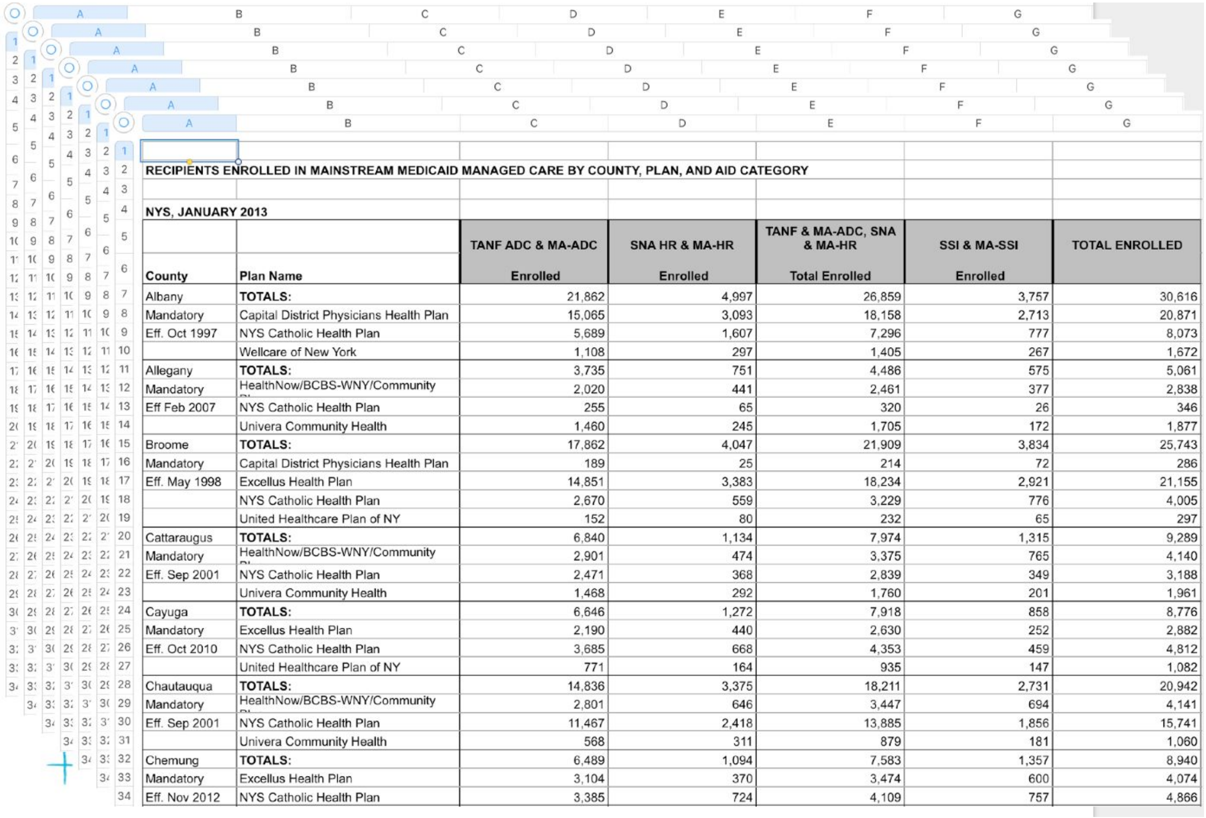
- Split data repeatedly on newline into rows
- Split data repeatedly on ','

(Screenshot Data Wrangler, Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

Another useful data cleaning tool is Data Wrangler. Designed by a data visualization team at the University of Stanford, it is especially useful when handling small data sets (up to 2000 lines). Compared to Open Refine, simple cleaning with Data Wrangler can be achieved with only a few steps.

(Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

Example 4: Advanced Cleaning with Python



RECIPIENTS ENROLLED IN MAINSTREAM MEDICAID MANAGED CARE BY COUNTY, PLAN, AND AID CATEGORY

NYS, JANUARY 2013

		TANF ADC & MA-ADC	SNA HR & MA-HR	TANF & MA-ADC, SNA & MA-HR	SSI & MA-SSI	TOTAL ENROLLED
County	Plan Name	Enrolled	Enrolled	Total Enrolled	Enrolled	
Albany	TOTALS:	21,862	4,997	26,859	3,757	30,616
Mandatory	Capital District Physicians Health Plan	15,065	3,093	18,158	2,713	20,871
Eff. Oct 1997	NYS Catholic Health Plan	5,689	1,607	7,296	777	8,073
	Wellcare of New York	1,108	297	1,405	267	1,672
Allegany	TOTALS:	3,735	751	4,486	575	5,061
Mandatory	HealthNow/BCBS-WNY/Community	2,020	441	2,461	377	2,838
Eff Feb 2007	NYS Catholic Health Plan	255	65	320	26	346
	Univera Community Health	1,460	245	1,705	172	1,877
Broome	TOTALS:	17,862	4,047	21,909	3,834	25,743
Mandatory	Capital District Physicians Health Plan	189	25	214	72	286
Eff. May 1998	Excellus Health Plan	14,851	3,383	18,234	2,921	21,155
	NYS Catholic Health Plan	2,670	559	3,229	776	4,005
	United Healthcare Plan of NY	152	80	232	65	297
Cattaraugus	TOTALS:	6,840	1,134	7,974	1,315	9,289
Mandatory	HealthNow/BCBS-WNY/Community	2,901	474	3,375	765	4,140
Eff. Sep 2001	NYS Catholic Health Plan	2,471	368	2,839	349	3,188
	Univera Community Health	1,468	292	1,760	201	1,961
Cayuga	TOTALS:	6,646	1,272	7,918	858	8,776
Mandatory	Excellus Health Plan	2,190	440	2,630	252	2,882
Eff. Oct 2010	NYS Catholic Health Plan	3,685	668	4,353	459	4,812
	United Healthcare Plan of NY	771	164	935	147	1,082
Chautauqua	TOTALS:	14,836	3,375	18,211	2,731	20,942
Mandatory	HealthNow/BCBS-WNY/Community	2,801	646	3,447	694	4,141
Eff. Sep 2001	NYS Catholic Health Plan	11,467	2,418	13,885	1,856	15,741
	Univera Community Health	568	311	879	181	1,060
Chemung	TOTALS:	6,489	1,094	7,583	1,357	8,940
Mandatory	Excellus Health Plan	3,104	370	3,474	600	4,074
Eff. Nov 2012	NYS Catholic Health Plan	3,385	724	4,109	757	4,866

(Screenshot presentation Advanced Data Cleaning with Python, Tulio Pires 2016)

Marco Tulio Pires' School of Data presentation on Advanced Cleaning with Python 2.7 covered CSV and XLS file formats, common expressions and working with Python modules/libraries as well as the creation of algorithms for assigning values to variables, for if/else tests and for control flows. The presentation essentially aimed to provide help for any situation in which there are a number of files with the same structure that need to be consolidated into a single CSV file to aid further analysis.

Over the course of the session, participants were taken through one practical example using data from the Medicaid Managed Care Enrollment Reports. Pires' approach drew on Sarah Cohen's presentation on Advanced Data Cleaning with OpenRefine at NICAR 2016 and had been further developed whilst working on a script for Journalism++ Sao Paulo, a sister company to OpenBudget.eu consortium member Journalism++.

Pires encouraged the workshop participants to envisage the layout of the final document throughout the cleaning process by paying attention to the columns and what they individually represent and through looking for structural patterns between the files to be consolidated. In the workshop example, column headings included 'Year', 'Month', 'County' and 'Plan Name' and a number of inter-file patterns emerged. The dates (i.e. the values in 'Year ' and 'Month') were always in the same cell (A4) and the names of counties and plans were always to be found in the same columns in each file (columns A and B). After taking the participants through the cleaning code in Python, Pires concluded by pointing out that there will always be many ways to clean a file with Python; each user needs to work out his or her own approach to the task.[10]

To sum up, if you need to get rid of typos, blank spaces, duplicates, and reconcile different spellings, use:

- Open Refine

To change the type of your data or group/cluster the data, use:

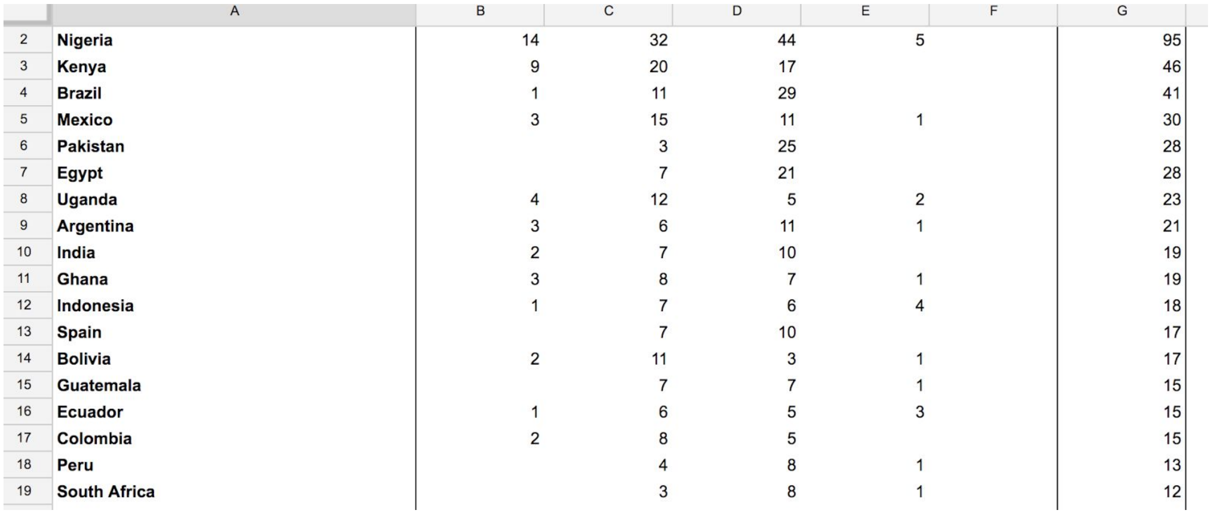
- Google Sheets (and alike)
- Data Wrangler
- Any programming language like Python

(Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

1.3 Tools to Analyze the Data

The next step in our workflow is the analysis of the cleaned data set to find angles and stories or just to make sure the data is consistent.

Example 1: Working with Google Sheets



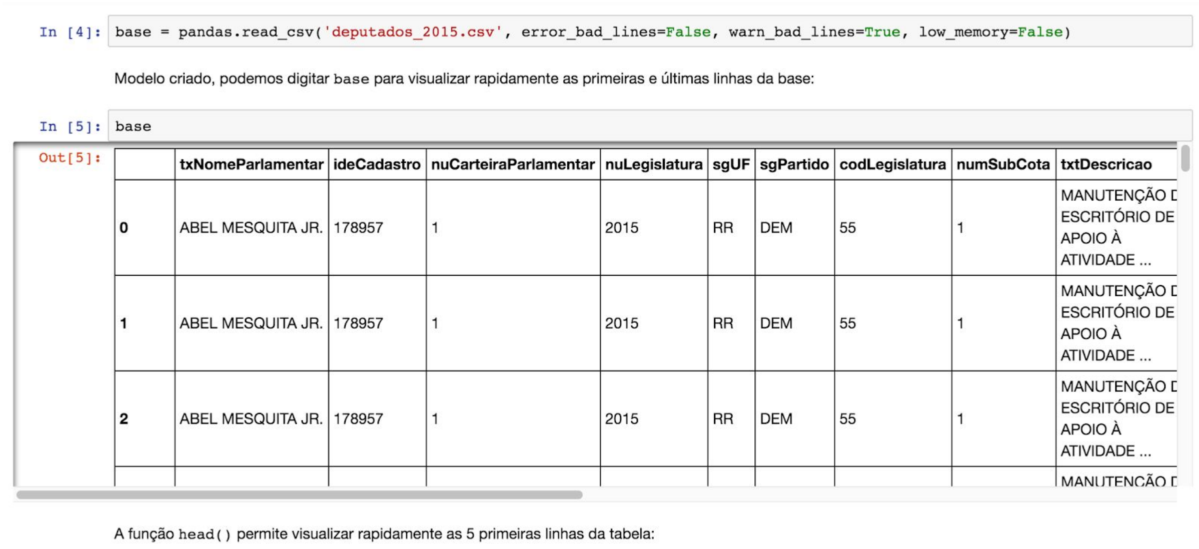
	A	B	C	D	E	F	G
2	Nigeria	14	32	44	5		95
3	Kenya	9	20	17			46
4	Brazil	1	11	29			41
5	Mexico	3	15	11	1		30
6	Pakistan		3	25			28
7	Egypt		7	21			28
8	Uganda	4	12	5	2		23
9	Argentina	3	6	11	1		21
10	India	2	7	10			19
11	Ghana	3	8	7	1		19
12	Indonesia	1	7	6	4		18
13	Spain		7	10			17
14	Bolivia	2	11	3	1		17
15	Guatemala		7	7	1		15
16	Ecuador	1	6	5	3		15
17	Colombia	2	8	5			15
18	Peru		4	8	1		13
19	South Africa		3	8	1		12

(Screenshot google sheets, Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016)

[10] The full code used by Pires at the workshop can be found here: https://github.com/mtrpires/pyPerugia16.

Many journalists are used to working with Pivot tables, for example, in Google Sheets. Pivot tables are used to consolidate data and eliminate the need to sum, add, subtract or calculate mediums and averages by hand. Pivot tables can aggregate data and show the reports that are useful and meaningful for analysis (Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

Example 2: Working with Python Pandas



```
In [4]: base = pandas.read_csv('deputados_2015.csv', error_bad_lines=False, warn_bad_lines=True, low_memory=False)
```

Modelo criado, podemos digitar `base` para visualizar rapidamente as primeiras e últimas linhas da base:

```
In [5]: base
```

Out[5]:

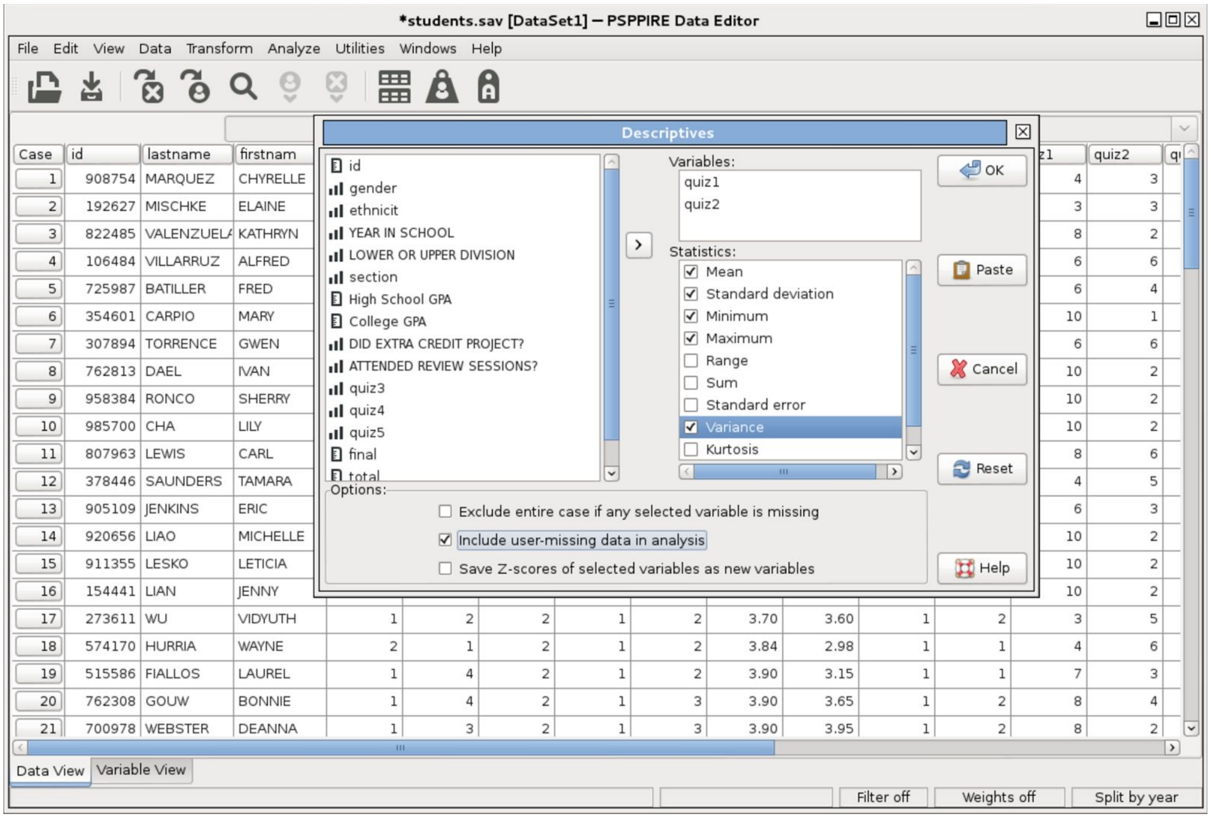
	txNomeParlamentar	ideCadastro	nuCarteiraParlamentar	nuLegislatura	sgUF	sgPartido	codLegislatura	numSubCota	txtDescricao
0	ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO D ESCRITÓRIO DE APOIO À ATIVIDADE ...
1	ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO D ESCRITÓRIO DE APOIO À ATIVIDADE ...
2	ABEL MESQUITA JR.	178957	1	2015	RR	DEM	55	1	MANUTENÇÃO D ESCRITÓRIO DE APOIO À ATIVIDADE ...
									MANUTENÇÃO D

A função `head()` permite visualizar rapidamente as 5 primeiras linhas da tabela:

(Screenshot Python Pandas - Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

Advanced data journalists can also use Python Pandas for statistical data analysis. Python is one of the most common programming languages used to analyse data, adapted by academia and also by newsrooms that have established data-driven journalism teams. Python Pandas is a convenient tool for handling big data sets that are too large for other tools to cope with (Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

Example 3: Working with PSSP



id
gender
ethnicit
YEAR IN SCHOOL
LOWER OR UPPER DIVISION
section
High School GPA
College GPA
DID EXTRA CREDIT PROJECT?
ATTENDED REVIEW SESSIONS?
quiz3
quiz4
quiz5
final
total

Variables:
quiz1
quiz2

Statistics:
☑ Mean
☑ Standard deviation
☑ Minimum
☑ Maximum
☐ Range
☐ Sum
☐ Standard error
☑ Variance
☐ Kurtosis

Options:
☐ Exclude entire case if any selected variable is missing
☑ Include user-missing data in analysis
☐ Save Z-scores of selected variables as new variables

Case	id	lastname	firstnam										z1	quiz2
1	908754	MARQUEZ	CHYRELLE										4	3
2	192627	MISCHKE	ELAINE										3	3
3	822485	VALENZUELA	KATHRYN										8	2
4	106484	VILLARRUZ	ALFRED										6	6
5	725987	BATILLER	FRED										6	4
6	354601	CARPIO	MARY										10	1
7	307894	TORRENCE	GWEN										6	6
8	762813	DAEL	IVAN										10	2
9	958384	RONCO	SHERRY										10	2
10	985700	CHA	LILY										10	2
11	807963	LEWIS	CARL										8	6
12	378446	SAUNDERS	TAMARA										4	5
13	905109	JENKINS	ERIC										6	3
14	920656	LIAO	MICHELLE										10	2
15	911355	LESKO	LETICIA										10	2
16	154441	LIAN	JENNY										10	2
17	273611	WU	VIDYUTH	1	2	2	1	2	3.70	3.60	1	2	3	5
18	574170	HURRIA	WAYNE	2	1	2	1	2	3.84	2.98	1	1	4	6
19	515586	FIALLOS	LAUREL	1	4	2	1	2	3.90	3.15	1	1	7	3
20	762308	GOUW	BONNIE	1	4	2	1	3	3.90	3.65	1	2	8	4
21	700978	WEBSTER	DEANNA	1	3	2	1	3	3.90	3.95	1	2	8	2

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

For those journalists who work with statistics, PSPP, the free version of SPSS, is another useful tool.

To sum up, if a visual overview of a small data set is required, use:

- Google Sheets

When handling big data sets and statistical analysis is needed to make sense of the data, use:

- Python Pandas
- PSPP

(Free Data Journalism Tools Slide Deck, Bouyer and Tulio Pires 2016).

1.4 Tools to Visualize the Data

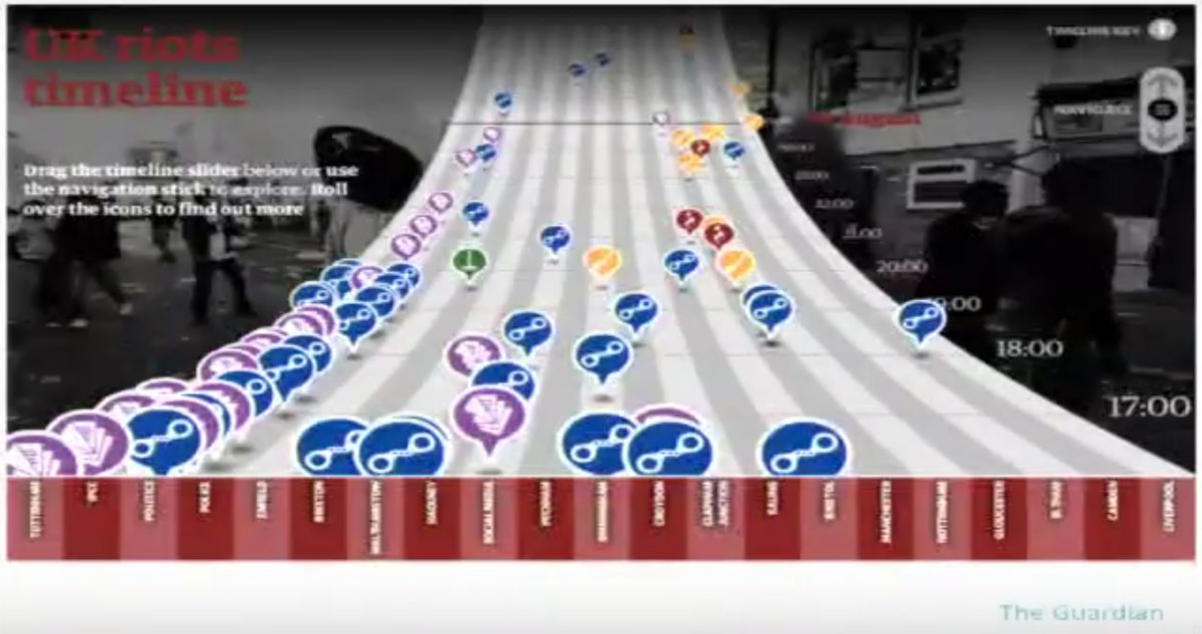
In her talk “Tips and Tricks for Data-Driven Journalism Starters”[11] at the International Journalism Festival 2016 in Perugia, Bahareh Heravi distinguished four different types of data visualizations: temporal (time, when), geospatial (locations, where), topical (text, what) and network (relationship, who) data visualizations.

Temporal data visualizations explore the question “when?”. Temporal data helps to understand temporal distributions of data sets, to identify growth rate, latency to peak times, or decay rates, as well as to locate patterns in time-series data, such as seasonality or bursts.

[11] Bahareh Heravi’s talk “Tips and Tricks for Data-Driven Journalism Starters” can be found here:http://www.journalismfestival.com/programme/2016/tips-and-tricks-for-data-driven-journalism-starters

Example 1: The Guardian, UK Riots Timeline

http://www.theguardian.com/uk/interactive/2011/sep/05/england-riots-timeline-interactive



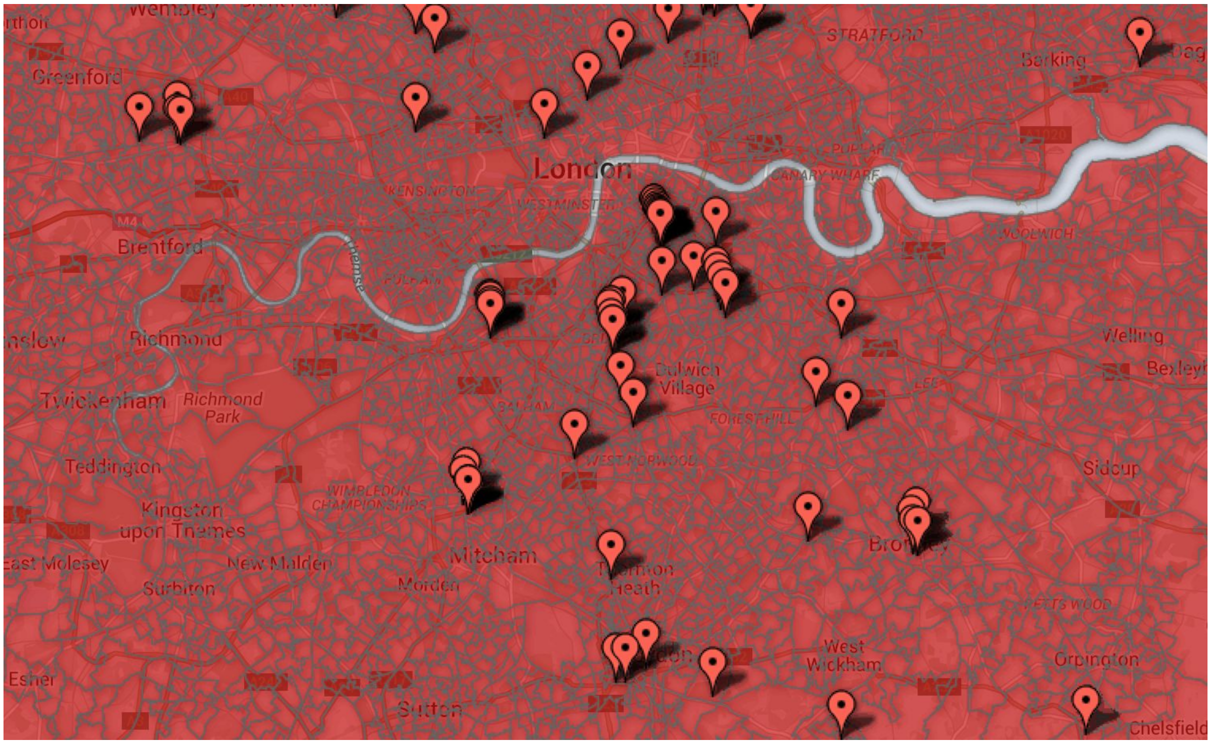
UK riots timeline

Drag the timeline slider below or use the navigation stick to explore. Roll over the icons to find out more

(Screenshot from Presentation Bahareh Heravi Perugia 2016)

Geospatial data visualizations explore the question “where?”. Geospatial data includes location information to identify positions, movements, trends or patterns (like tweets) over geographical space.

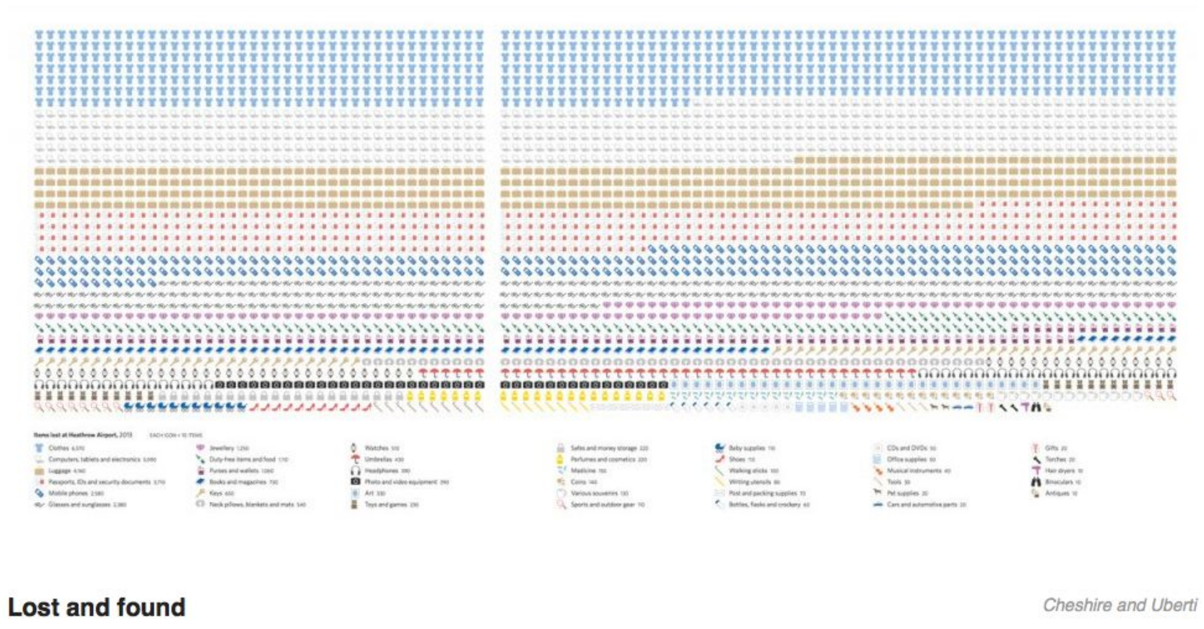
Example 2: The Guardian, Mapping the riots with poverty



Source:http://www.theguardian.com/news/datablog/interactive/2011/aug/10/poverty-riots-mapped (Screenshot from Presentation Bahareh Heravi Perugia 2016)

Topical data visualizations explores the question “what?”. Topical data includes text to identify major topics, their interrelations, and their evolution over time. The text can be a book, an article, a whole archive, or even just a tweet.

Example 3: Businessinsider, Lost & Found in Heathrow Airport



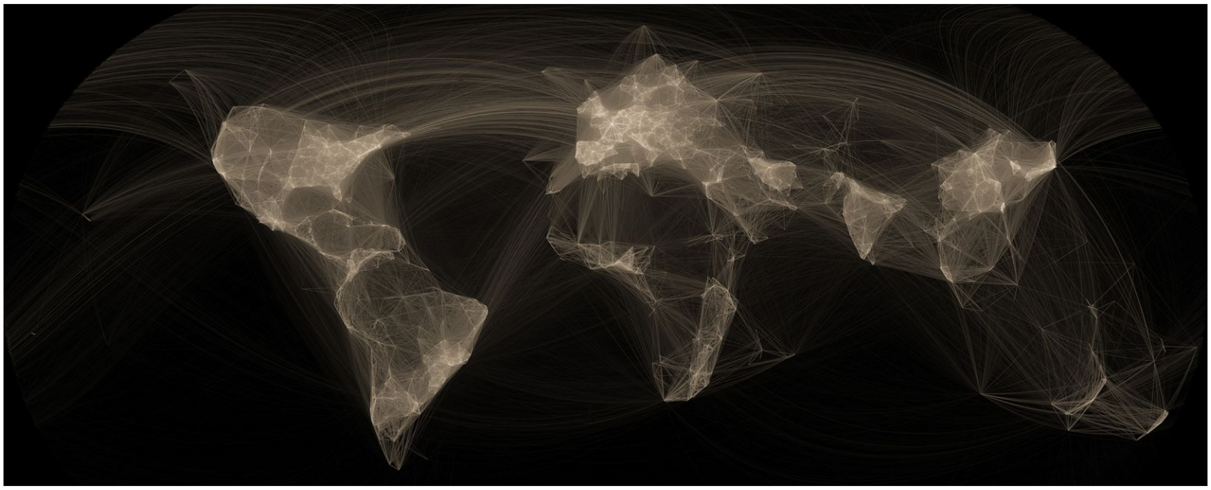
Lost and found

Cheshire and Uberti

http://www.businessinsider.com/maps-londons-bus-journeys-2014-11?IR=T (Screenshot from Presentation Bahareh Heravi Perugia 2016)

Network data visualizations explore the question “who?”. Network data identifies (highly) connected entities and the connections between them, network properties, such as size and density, and structures, such as clusters and backbones.

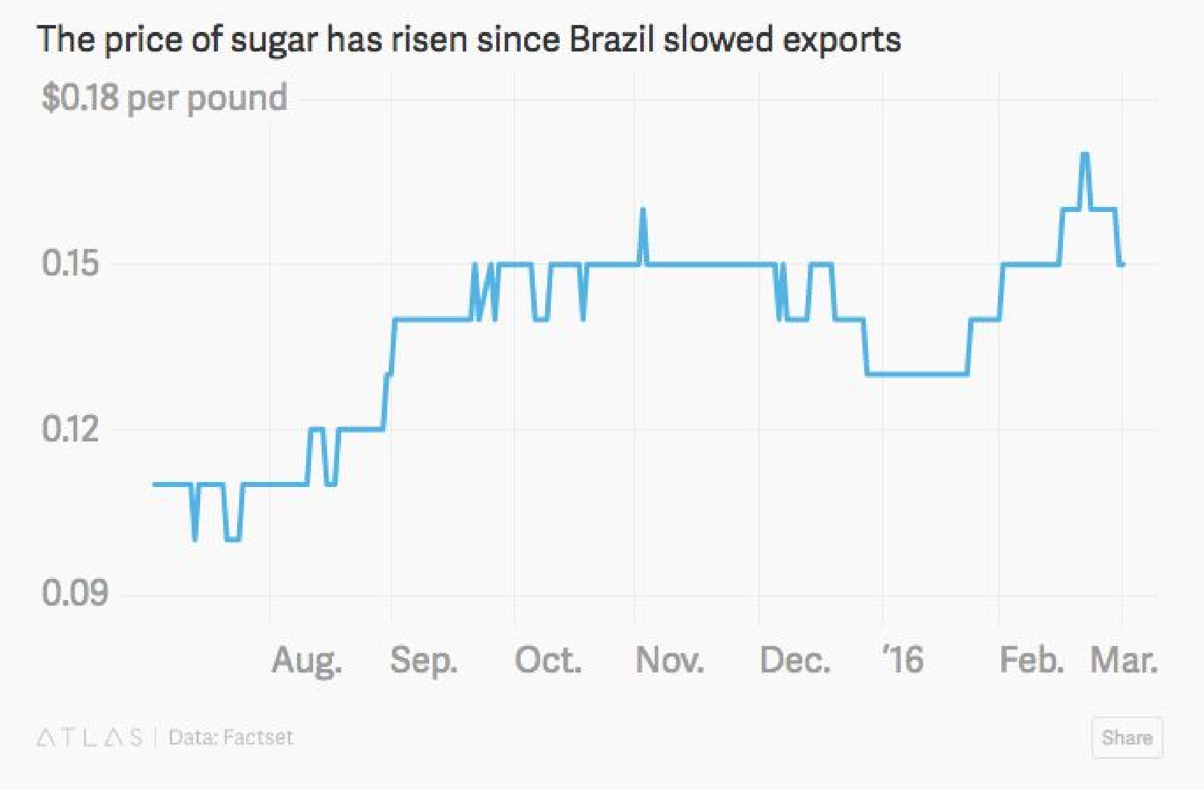
Example 4: Oliver H. Beauchesne, Map of Science Collaboration



http://olihb.com/2014/08/11/map-of-scientific-collaboration-redux/ (Screenshot from Presentation Bahareh Heravi Perugia 2016)

There are easy-to-use charting tools which allow for the creation of visualizations in just a few clicks, enabling reporters and editors to become more responsible for their own content and less dependent on those with specialized graphics skills. Visualizations also help readers to understand the data more efficiently.

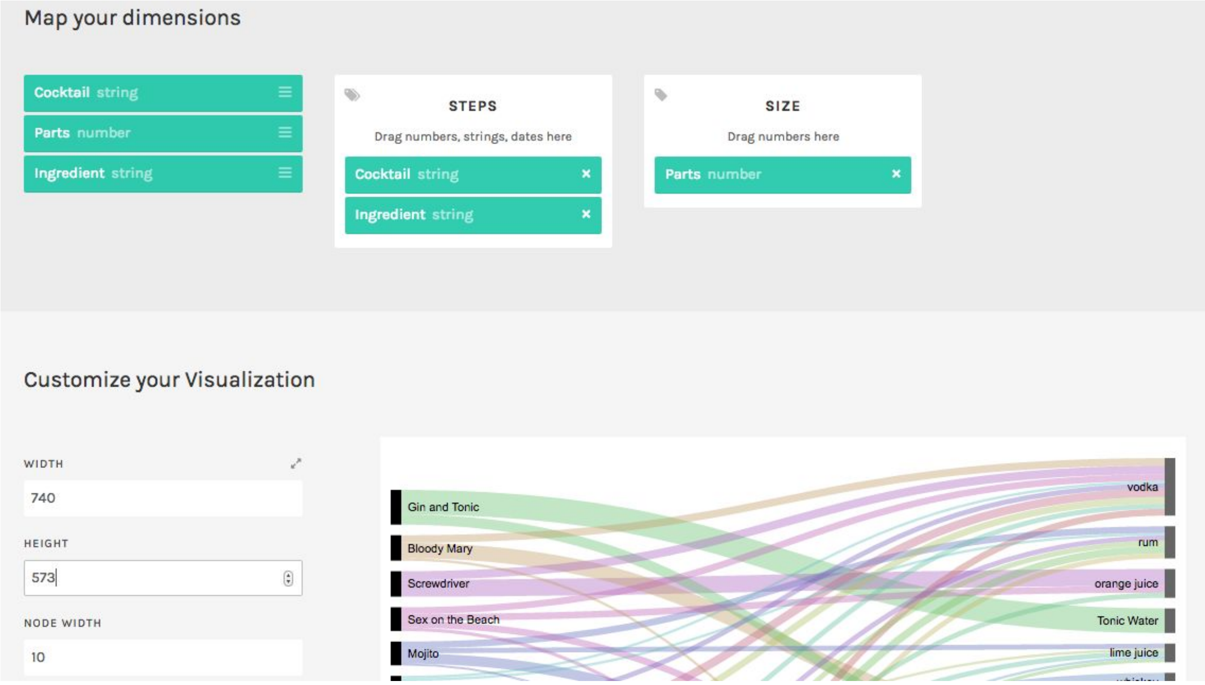
Example 5: Working with Chartbuilder



(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

One such simple tool for the creation of simple visualizations is Chartbuilder. It generates appropriate graphics based on the data set, which can be pasted directly into Chartbuilder from an Excel or csv file (Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016).

Example 6: Working with RAW



Map your dimensions

Cocktail string
Parts number
Ingredient string

STEPS
Drag numbers, strings, dates here
Cocktail string
Ingredient string

SIZE
Drag numbers here
Parts number

Customize your Visualization

WIDTH
740

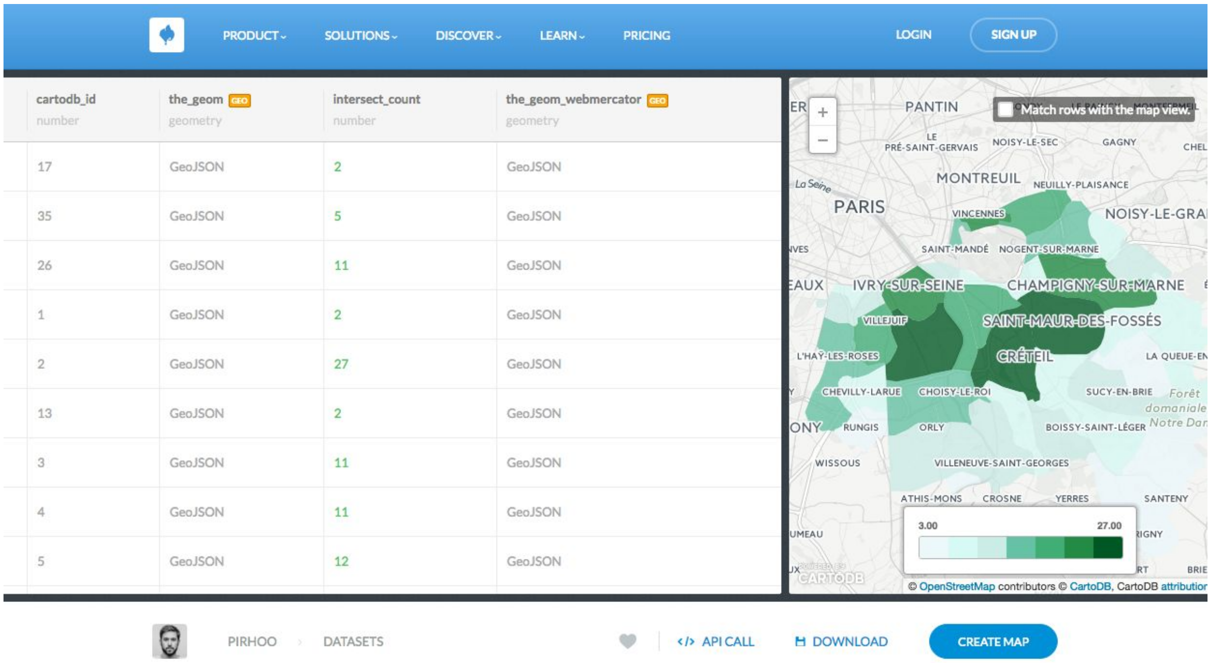
HEIGHT
573

NODE WIDTH
10

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

For complex visualizations, like alluvial diagrams, use RAW. RAW does not require any programming skills (Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016).

Example 7: Working with CartoDB



cartodb_id number	the_geom GEO geometry	intersect_count number	the_geom_webmercator GEO geometry
17	GeoJSON	2	GeoJSON
35	GeoJSON	5	GeoJSON
26	GeoJSON	11	GeoJSON
1	GeoJSON	2	GeoJSON
2	GeoJSON	27	GeoJSON
13	GeoJSON	2	GeoJSON
3	GeoJSON	11	GeoJSON
4	GeoJSON	11	GeoJSON
5	GeoJSON	12	GeoJSON

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

A commonly-used tool for producing maps is CartoDB. Like RAW, CartoDB does not require any programming skills.

To sum up, if journalists need to create simple visualisation for readers or simply to understand your data, use:

- Chartbuilder

To create complex visualization or maps without using any codings skills, use:

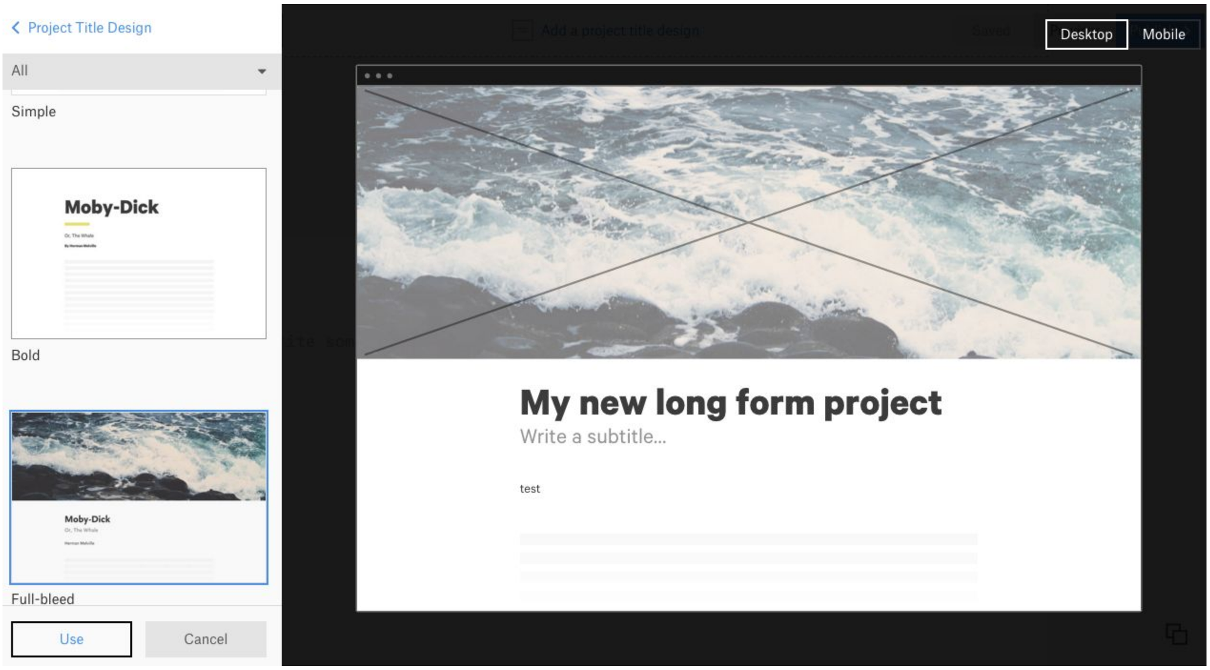
- CartoDB
- RAW

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

1.5 Tools to Tell a Story with Data

Finally, journalists need to create a story that communicates the most relevant information to the reader. This last section covers the final step of the data pipeline, offering easy-to-use tools to help tell a compelling story.

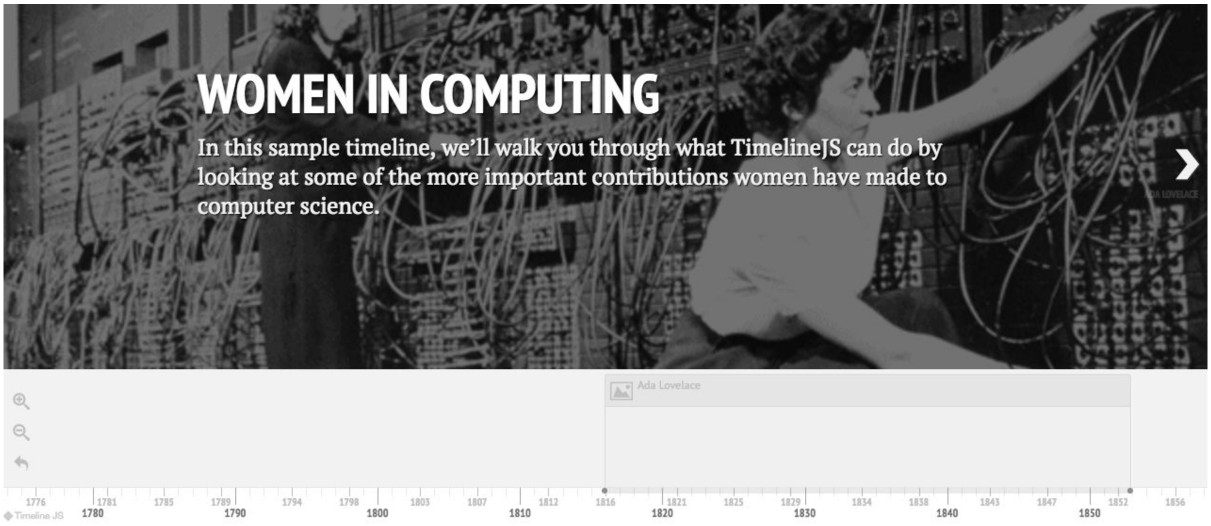
Example 1: Working with Atavist



(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

With Atavist, journalists can make, design and share their story just by dragging different media types around (Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016).

Example 2: Working with timeline.js



WOMEN IN COMPUTING

In this sample timeline, we'll walk you through what TimelineJS can do by looking at some of the more important contributions women have made to computer science.

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

You can use timeline.js to tell a story in a chronological order (Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016).

Example 3: Working with Silk.co

Silk lets anyone create interactive data visualizations, publish websites, and tell interactive stories.

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

With Silk.co, journalists can make interactive data visualizations, publish websites and produce interactive stories.

To sum up, if you want to tell a story in an chronological order, use:

- Timeline.js

If you need an already prepared template for your story, use:

- Atavist

To create a story with interactive visualizations, use:

- Silk.co

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016).

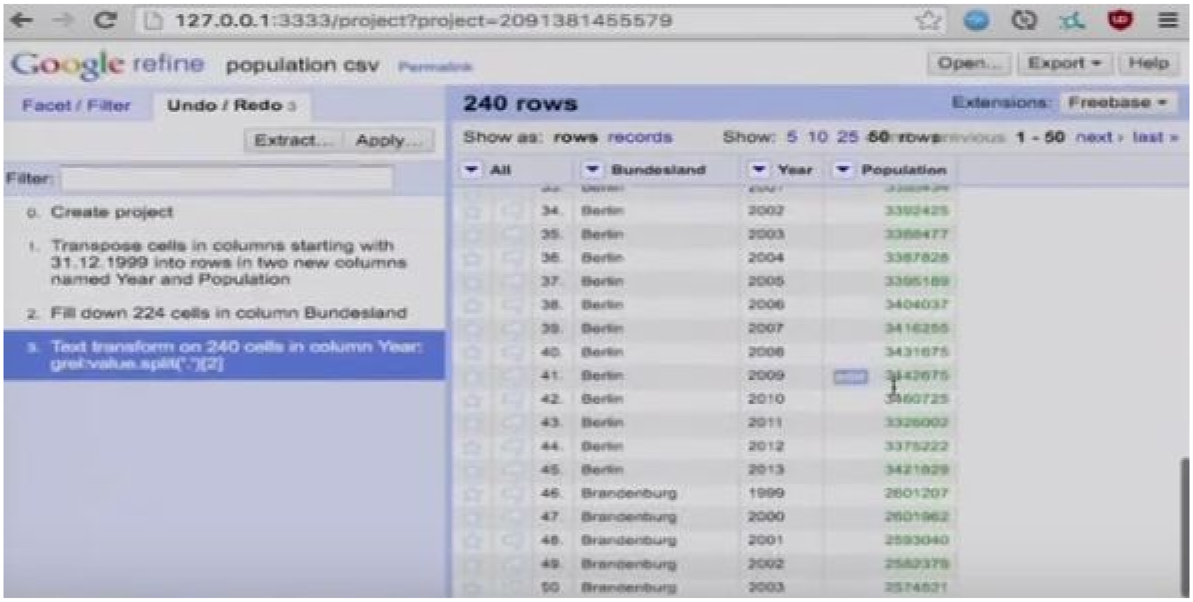
1.6 Tools to Ensure a Reproducible and Transparent Data-Driven Journalism

Aiming for transparent data-driven journalism is not only a matter of demonstrating integrity and boosting credibility, a higher degree of transparency can simply make stories more understandable for readers. A transparent working method that details the data source and the processing steps used is particularly important when dealing with large datasets where complete accuracy is very difficult to ensure.

As many different interpretations of the same dataset are usually possible, it is essential that the reader can follow the process of how a particular interpretation was reached and which method of analysis was used. Ideally, this transparency should make it possible for someone else to reproduce the interpretation and reach the same conclusion; it should be a replicable result.[12]

[12] Stefan Wehrmeyer's talk "Reproducible and transparent data journlism with modern tools" held at the International Journalism Festival 2016 in Perugia is available online: http://www.journalismfestival.com/programme/2016/reproducible-and-transparent-datajournalism-with-modern-tools.

Example 1: Working with Open Refine



127.0.0.1:3333/project?project=2091381455579

Google refine population csv Permalink

Open... Export ▾ Help

Facet / Filter Undo / Redo 3

Extract... Apply...

Filter:

0. Create project
1. Transpose cells in columns starting with 31.12.1999 into rows in two new columns named Year and Population
2. Fill down 224 cells in column Bundesland
3. Text transform on 240 cells in column Year: grel:value.split('.')[2]

240 rows

Extensions: Freebase ▾

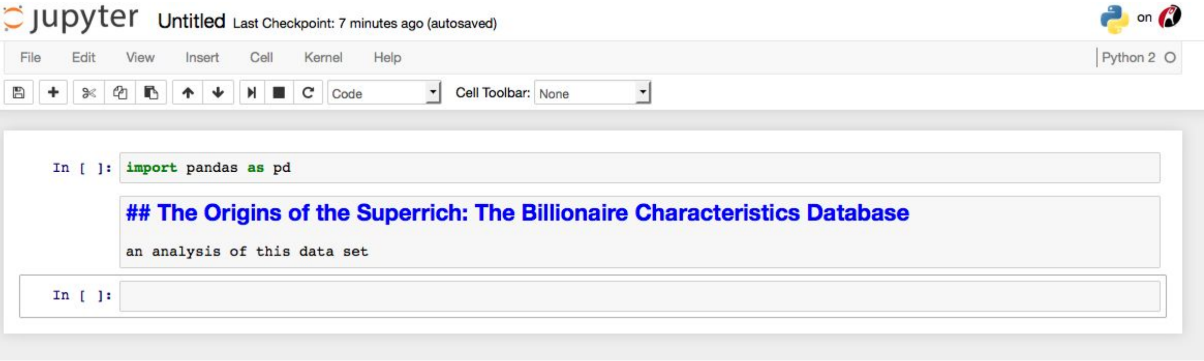
Show as: rows records Show: 5 10 25 50 rows « previous 1 - 50 next › last »

All	Bundesland	Year	Population
34.	Berlin	2002	3392425
35.	Berlin	2003	3388477
36.	Berlin	2004	3387828
37.	Berlin	2005	3395189
38.	Berlin	2006	3404037
39.	Berlin	2007	3416255
40.	Berlin	2008	3431675
41.	Berlin	2009	3442675
42.	Berlin	2010	3460725
43.	Berlin	2011	3326002
44.	Berlin	2012	3375222
45.	Berlin	2013	3421829
46.	Brandenburg	1999	2601207
47.	Brandenburg	2000	2601962
48.	Brandenburg	2001	2593040
49.	Brandenburg	2002	2582379
50.	Brandenburg	2003	2574521

(Wehrmeyer Perugia 2016)

Open Refine is useful for many steps in the data pipeline. This tool can also help to ensure the transparency and reproducibility of analysis methods as all steps of data cleaning are recorded (undo/redo section). By using the extraction function, the user can create both a machine readable and comprehensible version, displaying all steps that were taken. In this way, code can be shared and easily updated at the same time.

Example 2: Working with Jupyter Notebooks



```
In [ ]: import pandas as pd
```

The Origins of the Superrich: The Billionaire Characteristics Database

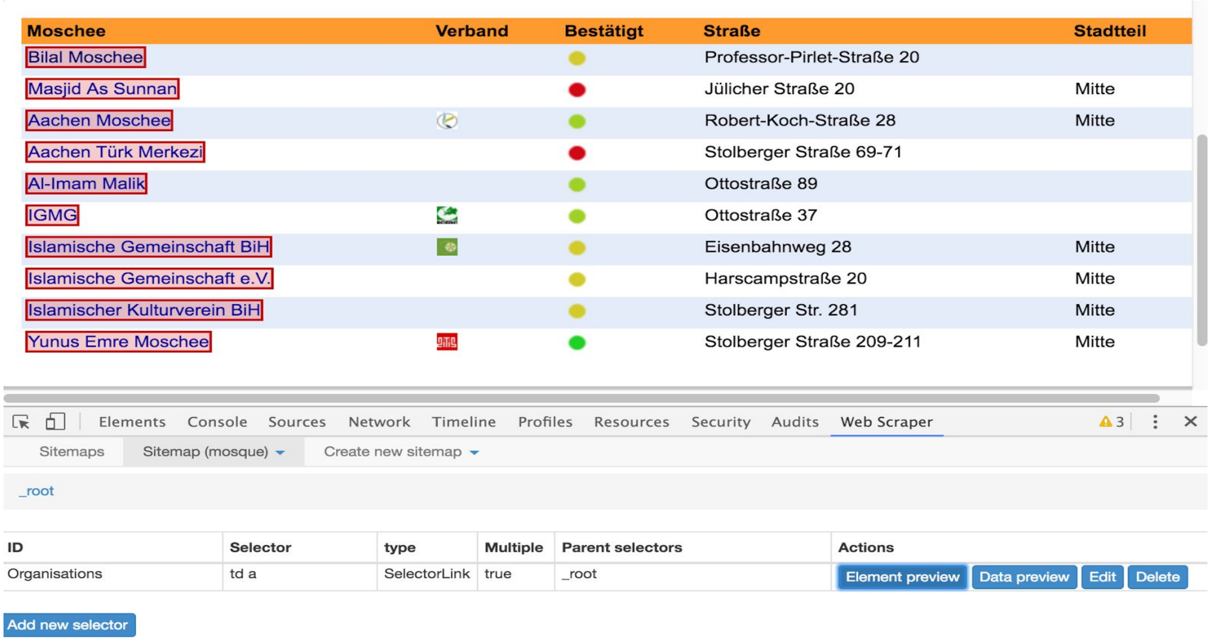
an analysis of this data set

```
In [ ]:
```

(Wehrmeyer Perugia 2016)

When working with a complicated data set, Jupyter Notebooks is another commonly used solution for ensuring transparency of workflows and providing a documentation. Jupyter is an environment that runs directly in a web browser. It provides a web interface or “notebook” that not only shows data and code but also the result of the code execution. Jupyter Notebook also provides many ready-to-use libraries (such as Python Pandas, mentioned above), that can be used for different types of data analysis and workflow documentation.

Example 2: Working with Web Scraper



Moschee	Verband	Bestätigt	Straße	Stadtteil
Bilal Moschee		🟡	Professor-Pirlet-Straße 20	
Masjid As Sunnan		🔴	Jülicher Straße 20	Mitte
Aachen Moschee		🟢	Robert-Koch-Straße 28	Mitte
Aachen Türk Merkezi		🔴	Stolberger Straße 69-71	
Al-Imam Malik		🟢	Ottostraße 89	
IGMG		🟢	Ottostraße 37	
Islamische Gemeinschaft BiH		🟡	Eisenbahnweg 28	Mitte
Islamische Gemeinschaft e.V.		🟡	Harscampstraße 20	Mitte
Islamischer Kulturverein BiH		🟡	Stolberger Str. 281	Mitte
Yunus Emre Moschee		🟢	Stolberger Straße 209-211	Mitte

ID	Selector	type	Multiple	Parent selectors	Actions
Organisations	td a	SelectorLink	true	_root	Element preview Data preview Edit Delete

(Free Data Journalism Tools Slide Deck - Bouyer and Tulio Pires 2016)

Another useful tool for obtaining data is Web Scraper, which is a browser extension for Google Chrome. It’s a point and click tool that allows you to easily select elements on a web page and then extract the information into a data set.

Ensuring Quality Trainings

This box covers the most relevant aspects that need to be considered for a successful workshop:

Before the training

- Provide a detailed description of your training to target the most suitable audience;
- Let your audience know if your trainings are suitable for beginners, intermediates or experts;
- Prepare all data sets in an appropriate way for a specific exercise so that your audience can easily understand and work with your data;
- If you are offering a participatory data training, make sure to provide a detailed documentation (like a Github repository) beforehand, where each step is explained and can be easily followed. Include all data sets you are working with in the documentation;
- Tell your participants what they should bring to the training (laptops, hardware) and what will be provided;
- Let your participants know what they need to prepare before the training (e.g. any data/software that need to be downloaded).

During the training

- Explain the topic, make sure your audience understands why the aspects you are addressing are relevant for them or wider society, communicate the goal of the training;
- Always start with an interesting questions to capture attention;
- Always keep in mind: your training is about a topic, a question or/and a problem you want to solve. Technology and data can help to find the solution and to tackle the problem, but they are not the main focus of discussion.;
- To ensure that all non tech-savvy participants feel involved, you should communicate clearly and use a language that everyone understands (avoid using too many technical terms, use examples, etc.)
- If you show code, explain the code;
- Leave enough time for questions;
- Never touch the laptops of the participants or directly write code for them, try instead to guide them towards the solution.

After the training

- Ask for feedback on your training;
- Make sure your documentation is updated and available online;
- Stay in touch;
- Provide further information on the topic.

Box 1: do’s and dont’s for training and workshops