OpenBudgets.eu

OpenBudgets.eu: Fighting Corruption with Fiscal Transparency

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

Deliverable 5.1

User Requirement Reports First Cycle

Dissemination Level	Public
Due Date of Deliverable	Month 8, 30.12.2015
Actual Submission Date	19.12.2015
Work Package	WP 5, Test Beds and Evaluation - Journalism
Task	T 5.1
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	27
Filename	D5.1 User Requirement Reports - First Cycle.docx

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

Project funded by the European Union's Horizon 2020 Research and Innovation Programme (2014 – 2020)

History

Version	Date	Reason	Revised by
0.1	26.11.2015	First version	Dominik Wurnig
0.2	03.12.2015	Anonymized the text	Nicolas Kayser-Bril
1.0	17.12.2015	Final version	Nicolas Kayser-Bril

Author List

Organisation	Name	Contact Information
OKFDE	Anna Alberts	anna.alberts@okfn.de
OKFDE	Dominik Wurnig	Dominik.Wurnig@okfn.de
J++	Nicolas Kayser-Bril	nkb@jplusplus.org
J++	Anne-Lise Bouyer	annelise@jplusplus.org

Executive Summary

The user requirement research and the conducted interviews revealed that journalists and more specifically investigative and data journalists have limited use of data portals like SpendingStories or similar tools. Leaders in the field criticized data portals for their lack of stories, raw data and exclusivity. From a journalistic perspective, budget data needs to be looked at from a different angle than data repositories. Analyzing real life case studies - The Migrants' Files and The Football Tax - we identified several gaps why journalists do not thoroughly report on budget data.

The gaps standing between journalists and budget data include formal barriers like unavailability, refusal to publish, format, timeliness, completeness, identification and standardization. Intended or unintended, those barriers prevent journalists from properly and easily accessing budget data and therefore cover it in-depth. To bridge those gaps, we offer.

Other gaps concern the lack of budget data and accounting expertise among some journalists, who lack the in depth knowledge of the public finance and administration system. They will be overcome by the workshops and tutorials part of this work package. As a result of the cases studies and the gap analysis we identified a need for journalists to be trained in accounting, budget forecasts, fiscal transparency, government spending and budget data together.

The gap analysis finally revealed that hardly any curricula among journalism schools targeted both data journalism and budget data expertise. A few offer specialization in one or the other but extremely rarely in both. This work package will focus on creating and establishing a training program and training material for journalists interested in learning how to collect, interpret and communicate public budget data.

Abbreviations and Acronyms

FTS	Financial transparency system of the European Commission
ICIJ	International Consortium of Investigative Journalists, a nonprofit based in the United States.
TFT	The Football Tax, a WP5 case study
TMF	The Migrants Files: Money Trails, a WP5 case study
WP	Work Package

Table of Contents

List of Figures

List of Tables

1 Introduction

Recent analysis of open data initiatives highlighted the lack of systematic assessment of current tools and platforms. Jonathan Gray of Open Knowledge, for instance, found that little attention is paid to the "demand side" of open budget data[1]. Rebecca Rumbul of MySociety notes that usage of civic tools is extremely rarely evaluated[2].

Participants of the OpenBudgets.eu consortium have similar experiences. A lot of energy has been put in finding data sources, cleaning data sets, analyzing them and presenting them in meaningful visualizations. These previous projects aimed at making data accessible in the hope of improving trust among stakeholders. In 2013, Journalism++ and Open Knowledge built SpendingStories, a tool that helped journalists make sense of orders of magnitude in spending data.

Despite very positive response from open data professionals, such tools are little used by their intended audience - journalists and citizens. Beyond the low usage numbers, they simply do not fit in the workflows of journalists.

From these observations, Work Package 5 is running a thorough gap analysis to precisely understand the needs of journalists who work to increase accountability in the public sector. It lets us develop appropriate tools and provide helpful feedback to the other work packages.

The gap analysis consists of case studies and interviews with high-profile data journalists and investigative journalists from the United States, France, Denmark, Romania, Germany, Spain and beyond.

1.1 The journalist's workflow

Previous projects have been built with assumptions about journalistic workflows that were not always verified. From our experience and the interviews we conducted, we examine how journalists go from data to story and vice-versa.

First of all, data always needs a story. Data by itself is dry, meaningless and uninteresting for the public. Presenting data without a story just for the sake of it or just to show a pretty visualization is considered to be superfluous by the interviewees. Data without interpretation and context is meaningless.

> *Even Wikileaks realized this in the beginning: We are dumping all this data, it is great data, why is nobody taking it. And that was not just databases, this was content. But still at first, people didn't pick it up. Then when Wikileaks recalled itself as media organization and started pushing packages and cooperating with journalists then it all started to take off properly.*
>
> – A Romanian journalist

Journalists usually start with a hunch or an idea for a story, which they go after. Reporting the whole story, investigative journalists keep on digging until they find tangible evidence for every assertion. They explore any kind of source to find evidence: open data portals, data dashboards, offline records etc. The kind of data source where they find the missing link is different in every investigation.

> *From project to project we use whatever data we need to get the story done. Whether it is simple sitting there on a government website or whether it is something we had to aggregate ourselves.*
>
> – An American journalist

[1] *Open Budget Data: Mapping the Landscape* GIFT, digital methods initiative, Open Knowledge, 2015, p. 39 ff.

[2] *Who benefits from civic technology?* MySociety, 2015

Secondly, exclusivity is important in journalism. Generally available data (like open data) is considered less interesting and less newsworthy because everyone else may have seen it before. Despite recent changes in the media landscape, media organizations are more focused than ever before on scoops and exclusive stories. “The fact that something gets leaked makes it newsworthy, while something is just sitting there available it is routine, it is not newsworthy,” said Gebeloff. If a secret dataset gets leaked journalists make an exception and start digging in the data set to find a story. The reasoning is that it is very likely to find something interesting or juicy in there. Or, as a French investigative journalist puts it: ”What I feel difficult with open data is that I am not going there and find a story. It is very unlikely that I have the time to go there and figure out the story because it is likely there won't be any story.”

A major exception to those rules are data portals focusing on different kind of datasets which raise interest just by itself. Only datasets that are interesting for the public can attract a large audience to have reach and impact.

> *There are some media in the US where around 60 to 70 percent of their traffic is actually going to their databases. Take Texas Tribune. Their records on state employees and salaries. Salaries is something that attracts huge attention because all employees want to check what they get, if their information is right, what their colleagues are getting and what their bosses are getting. So you have a huge attention there. And then people who use these authorities are also interested in their salaries. So this is a kind of data set which attracts a lot of attention and where you might be forced to publish a lot of stories to make it very useful.*
>
> – A Danish journalist

Another example where a data portal attracts a lot of clicks without producing stories is Mugshots by the Tampa Bay Times, where pictures and information on persons recently arrested by the police is presented. *Mugshots* and *Salaries* by the Texas Tribune are two examples showing that there is a human interest in some data sets.

Third and most importantly, several journalist told us they prefer raw data over processed data.

> *I prefer having raw data. I am not interested in visualizations or in analysis done by others; I wouldn't use analysis done by others. I don't like tons of stuff that tries to give me the analysis. I want to have the raw data, I want to be able to download it easily, I want to have an index to understand the spreadsheet.*
>
> – A French investigative journalist

Data journalists and investigative journalist usually have a high understanding of technical and mathematical issues. Data portals do not offer them much added value. Regular journalists who are producing stories everyday usually do not have the time and capacity to look through data portals to sniff for stories. Often, data portals simply publish data sets that have already been published elsewhere.

> *A lot of those data portals are not taking into consideration the content of the data. Sometimes people are focusing just on the format: pdf is bad, csv is good. But that is not the point. The point is do you have that sentence that changes people's life or do you have a phrase that says nothing. And I don't care if the sentence is set in stone, so you can't scrape it or if it is in some open data format. Ideally you have it in open data format but I am talking about the real life.*
>
> – A Romanian journalist

Our own experience and interviews with experts showed that budget data in general does not fit this description: Budget data is hard to understand and to digest. Nevertheless, budget data is important and relevant to the public: it is policy cast in numbers. Therefore it is the journalists’ responsibilities to understand, question and explain public budget and spending to the public. From a journalistic perspective, budget data needs to be looked at

from a different angle than data repositories. We therefore designed case studies that help us understand how budget data can be used by journalists and what hurdles lie in the way.

1.2 Presentation of our case studies

This Work Package is centered on case studies, which will, complemented with original research and interviews of practitioners and experts, enable us to precisely assess the needs and demands of journalists working with budget data. Each case studies is an ensemble of elements that aims at tackling a journalistic issue that can be solved using budget data.

As of Month 8 of the OpenBudgets.eu project, two case studies are in progress. They were developed according to three criteria. First, that the journalistic question is topical and attractive to practitioners, so that the WP can collect as much feedback as possible from journalists. Second, that the journalistic question accepts a clear, measurable and concrete answer, so that it is possible to assess the degree to which the question was answered. Third, the journalistic question must be answerable using budget data from several administrative sources within European institutions or Member States, at the supranational, national or local level.

1.2.1 Case study 1: The Migrants Files: Money Trails

The Migrants Files is an award-winning[3] investigation started and coordinated by Journalism++. It involved at its peak over 25 journalists from 16 European countries.

The research question is as follows: How much do the policies that prevent migrants, asylum seekers and refugees from coming to or staying in Europe cost European taxpayers? The overall project started in 2013 encompassed other research questions, such as the amounts paid by refugees and migrants to facilitators and the number of men and women who died en route to Europe. However, this WP was only interested in the first research question and worked with The Migrants' Files journalists in June and July 2015.

In the course of this case study, journalists tracked and found public spending and budgets in a variety of sources, from official budgets to parliamentary reports to leaked documents and freedom of information requests[4]. The visualization of the results required the development of a new type of visualization, an interactive "Billion-dollar-o-gram"[5] (the name of which comes from previous work by visualization expert David McCandless[6]). The tool will be detailed in the deliverable 5.5.

1.2.2 Case study 2: The Football Tax

In order to better understand the workings of local budgets, which are at the heart of OpenBudgets, a second case study was launched, centered around the following question: How much public money is given to professional football teams in the European Union? The question was chosen because it relates to a great variety of financing types (subsidies, stadium building etc.), it is relevant in every single member state and it requires data from every administrative layer.

---

[3] It received the Data Journalism Award 2014, the European Press Prize 2015 and the ÑH Prize 2015.

[4] The list of expenses is available online and will be versed to the OpenBudgets database.

[5] The tool is available on the OpenBudgets repository at https://github.com/openbudgets/billion-dollar-o-gram. The tool will be presented in more details in D5.5.

[6] His work can be found at http://www.informationisbeautiful.net/visualizations/billion-dollar-o-gram-2013/

1.3 Presentation of our workflow and methodology

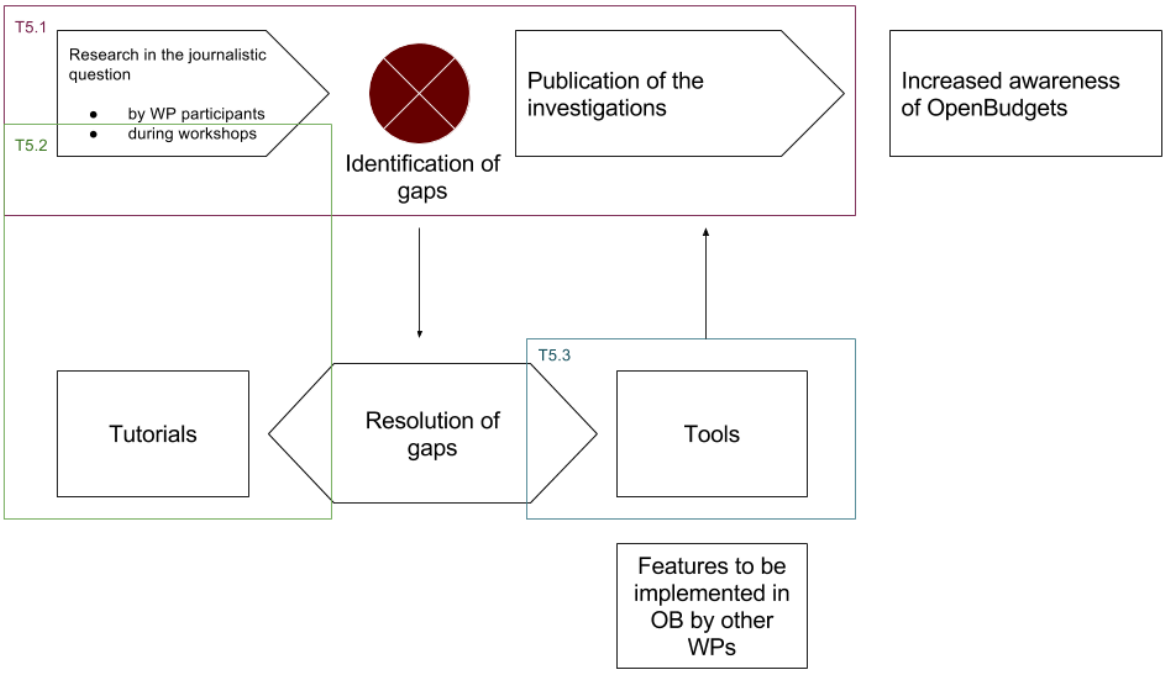

Figure 1 - Methodology of WP5

For each case study, we develop a workflow to answer the research question. In the course of this process, we uncover the points that block journalists in their work. These blocking points, in turn, let us complete the gap analysis, determine user requirements as solutions to these blockers and contribute to listing priorities for the overall OpenBudgets project.

We use a variety of techniques to answer the questions raised in a case study:

- Workshops and focus groups: Journalists and journalism students are tasked with or asked to reflect on how to answer a research question in order for us to observe journalists in their environment and see which problem they encounter and how their deal with budget data in their daily work.
- Interviews: We conducted several semi-directed interviews with thought leaders and practitioners in order to find the common problems and challenges that might not have appeared in the case study themselves but are nevertheless part of the user requirements.
- Direct data gathering: In order for us to be most efficient when organizing workshops and interacting with journalists, we set out to gather part of the data ourselves and in partnership with select journalists interested in publishing the outcome of the case study, using the following sources:
    - Open source intelligence (OSINT), which encompasses all available sources that have been already published, such as news archives, public statements by administration officials, deliberations of city councils, leaked documents etc.
    - Open data portals
    - On the ground research from partner journalists

2 Definitions of the gaps

> *We have to drop the belief that budgets are somehow neutral and self-explanatory objects and instead treat them as the weird mix of political process and creative accounting that they are. That doesn't mean that data is unimportant but we need to use it in ways which are less naive.*
>
> – A German journalist

For reasons detailed in this section, the use of budget data by journalists often lacks contextualization, as the German journalist, a leading figure in the field of data-driven investigations, puts it. This, in turns, prevents the realization of compelling journalistic work. This Work Package analyzed the gaps to be filled so that more journalists can make effective use of budget data.

Each identified gap is numbered to facilitate communication and identification of progress in the course of the overall project. The case studies that let us identify a gap are indicated in parentheses.

2.1 Access to budget data

2.1.1 G5.1 Unavailability

Both case studies showed that accessing a precise item of a budget or spending was never a trivial task, whether at the local, national or European level.

Most of the information needs to be extracted with ad-hoc means, either from the deliberations of the local public body (TFT in most French cities) or from ledgers (TMF made use of the European Commission financial transparency system). Some public bodies only store paper-based archive, such as the city of Sion, which has no budget data available in digital format prior to 2010 (TFT, Switzerland). Others do a poor job of keeping their online files accessible. In Neuchâtel (TFT, Switzerland), only one out of six financial report by the cantonal administration was available online. Links to all others offered only “404 - page not found” errors.

Interviews conducted with experienced investigative journalists show that accessing budget data is often obstructed by language barriers and varying publication standards in different regions and countries. A New-York-based journalist said that this is not just a European problem. In the United States, jurisdictions, laws and budget reporting differ tremendously throughout the states and counties. A Swiss journalist said that budget data for the Swiss context is difficult and hard to compare because of the cantonal system. There is one federal budgets and 26 cantonal budgets - each one is uses its own system.

A data journalist in Madrid gives this account:

> *In Spain, budget research is particularly hard because the regional and federal administration is very dispersed and messy. There are in total 70 different sources. Moreover, major posts such as health and education are switched from the federal to the municipal level. So it seems that the budget for these expenditures have diminished at one level, but in reality they have just been moved around. To get a picture of the entire budget on for example health care you need 17 or 18 searches.*

A Spanish investigative journalist pointed out in an interview that the situation is different in every country. In one country, education and health issues might be concerns of the central government, in other countries it might be decentralized. The explanation is not one in numbers or context, but a political explanation and analysis, she said. Therefore a data portal would also need to include political and contextual explainers.

2.1.2 G5.2 Refusal to publish

Some public bodies simply do not release the documents they are legally bound to publish, such as the decisions taken by the city council - which, in turn, contain some budget data (TFT, Nîmes, France). Some refuse access to data on national security grounds, as a French journalist found out when researching the amounts given to French companies as part of the Horizon2020 Security program. (The data was obtained via the EC instead).

As part of the case studies, several freedom of information requests have been filed. The next user requirement report will examine the success rate of such endeavours.

2.1.3 G5.3 Format

The case studies showed that the best format journalists can hope for are tables within PDF files (either text-based or scanned images) at the sub-national level and Excel files at the national level. (One notable exception being the FTS of the European Commission). The issue of format was highlighted by most interviewees.

The Spanish investigative journalist explains that web-scraping, a method that requires advanced computer skills, is often the only way forward:

> *Another difficulty is getting the data and the formats of open data. The different ministries use different data formats, mostly Excel. For others it is simply not available and they have to scrape the data from websites. In Spain the freedom of information act only applies to the national level, for regional and municipal governments they have to ask or scrape the data.*

2.1.4 G5.4 Timeliness

Case studies showed time and again that few data repositories were updated in time and in a consistent format. Even the FTS, by far the most useful data source among the open data repositories encountered by this WP, does not offer a consistent data format across years (TMF's study of research projects showed that amounts for individual beneficiaries were not available across all years from 2009 to 2013). Open data repositories rarely publish new data sets as the administration produces them. None of the data repositories journalists encountered in the case studies of this WP offered a planning of publication.

2.1.5 G5.5 Completeness and level of details

Published budget data is often detailed to the largest possible position, i.e it is aggregated to the point when it no longer is useful for journalists working on the case studies. Some cities publish very detailed data, but do not detail the contents of the accounting lines (TFT, Thessaloniki, Greece and TFT, Cologne, Germany). In such cases, journalists have access to all the accounts of the city, except to the finest level of details -- the accounts of the beneficiaries. The same is true of private bodies that have an obligation to publish their results, such as associations. Only a detailed view of the accounts can yield journalistically useful information (TFT, Nice, France).

The illustration below (excerpt from the books of the Association OGC Nice, France), shows a detailed account ("Subventions d'exploitation", operating subsidies, which is account 74 as per the French general accounting plan). The most interesting element for journalists, however, are the nominal accounts, such as 741, 742 etc., which are linked to a payer.

Subventions d'exploitation						
SUBVENTION EMPLOI	1 800	0,25	1 500	0,21	300	20,00
SUBVENTIONS MUNICIPALES	276 890	39,00	220 045	30,65	56 845	25,83
SUBVENTION CONSEIL REGION	98 000	13,80	263 500	36,70	-165 500	-62,81
SUBVENTION CONSEIL GENE	7 295	1,03	8 148	1,13	-853	-10,46
RECETTES TAXE APPRENTISSA	264 325	37,23	221 768	30,89	42 558	19,19
	648 310	91,32	714 960	99,58	-66 650	-9,32

Figure 2 - Illustration of a detailed account

2.1.6 G5.6 Identification of private beneficiaries

Third parties receiving public money can use several vehicles. Such vehicles can change over time, when a company that goes bankrupt and is replaced by the one that bought the assets of the former, for instance. In these cases, the public authorities can still have the former names in their financial systems, which adds a layer of ambiguity to the budgets (TFT, Strasbourg, France).

Additionally, it is often non-trivial to distinguish between companies of the same name that are actually distinct fiscal entities (TMF, example of Infoterra in France, Germany and the United Kingdom).

Finally, the beneficial ownership and the degree of independence enjoyed by subsidiaries is impossible to assess automatically. Only a case-by-case approach lets a journalist know if the amount given to a company is used by the nominal beneficiary or by its parent company (TMF, research on beneficiaries of FP7 projects).

2.1.7 G5.7 Identification of co-financing

The budgets of different administrative levels might count the same expenditure in the same budget line, even if an administration is only funding the other. Administration A might fund B to accomplish mission X. However, the amount budgeted by A might be found as a revenue line of administration B. In the absence of traceability, it is impossible to know precisely the total amount spent on mission X, all the more when administration B adds in-kind resources to the amount paid by A. This happens very often in projects co-financed by either the European Commission or national authorities (TMF, for every co-financed mission). It is for instance the case in Belgium, where the Wallonie Region gives subsidies to Provinces like Liège to support professional football (TFT, Liège, Belgium). Then Liège gives subsidies to specific clubs. It is impossible to find exactly which amount of money received the club from which public body.

Legal or contractual dispositions might also prevent the technical payer from having any agency over a budget item. If entity A gives money to entity B, which has, by law, to transfer it to entity C, entity B's role, for all intent and purposes, is void. In Switzerland, it can be the case that subsidies are channeled to three different bodies before reaching the one that actually takes the decision to allocate funds to a beneficiary (TFT, Lausanne, Switzerland). At the EU level, some funds are earmarked for certain member states without the Commission having any power as to whom or how the funds are spent (TMF, regarding the AMIF fund).

2.1.8 G5.8 Lack of standardization

In the private sector nearly all publicly trade companies and also many smaller corporations are following Generally Accepted Accounting Rules (GAAR). But the world of public budgets is different. There is no EU-wide standard for accounting and in federal countries like Switzerland there is not even a country-wide standard, as Sylke Grunewald, the Swiss journalist, points out. Moreover, the example of the Netherlands show that even when there is a highly standardized system for public budgeting, spending and accounting, municipalities still use their own accounting

system for concrete and detailed budget and expenditure data because the standard is too abstract for the day-to-day practice. This means there is only little standardization how certain expenditures should be accounted for. Without standardization it is next to impossible to compare budgets from different jurisdictions. Time comparisons in the same jurisdiction can also prove to be unreliable.

A journalist from an international consortium speaks about her experience as data journalist:

> *The same categories might change over time. For example under transportation the government understands in one year cars and buses, and the next year, cars, buses and trains. So they will change small parts of the budget; so on that level you need to understand the tweaks. You should not underestimate the classification and standardization and its influence on data. In a data portal you would have to indicate very clearly, when something is not comparable. For that the key is annotation on the level of the single value/observation.*

An experienced, American data journalist, agrees:

> *I think with something like budgets, the one thing everyone is afraid of is if you are trying to compare two countries' budgets or even within a country the budget of two different agencies: is it really standardized? Like, is the row that says personal costs the same cost in two agencies. Are they counting the same thing? Is it an apple to apple comparison?*

He suggests a very time intensive way to overcome this problem: aggregating your own data.

> *When I worked for a regional newspaper in New Jersey, we had an initiative called Standardized Town Reporting. New Jersey has 566 municipalities and as a goal we wanted to an ongoing series of stories that compared these communities in different metrics. And whenever we tried to do something budget related we discovered that towns often didn't report the budgets in similar ways and then we had to resort to our own surveys. Like we wanted to know what our police costs we would send our own survey to the town and say this is our definition of police costs. You know the salaries of police officers, the benefits of police offices, the salaries of other employees and so on.*

A New-York-based data journalist gives an example how his publication is overcoming the lack of standardization. In their investigations, they often assembles datasets through scraping. For example, they did a story on pharmaceutical companies. The information was only disclosed by some companies and in very different formats. “In the analysis process, we had to built a standard ourselves,” he said. “We needed that standard for the comparison, so if no one else does it, why not us.” If you make a standard, you have to be clear about the methodology, do it diligently and transparent, he said. And even if that standard in itself becomes topic of debate, he thinks it is all the better. That means people engage with the story and explain from their view why the standard is (not) applicable, he said.

Case studies showed that, in Europe, too, the same thing can be accounted for in very different ways. Cities can finance football clubs in two main ways. Either with a subsidy or through a sponsorship agreement, whereby clubs receive a lump sum in exchange for advertising and entry tickets to games (TFT, Liège, Belgium and Nice, France). In Liège, such expenses are accounted for as “entertainment expenses” because the contract lets the city invite guests to the VIP boxes in the stadium. In France, by contrast, such expenses fit in the account ‘023 - Communication and publicity’ (according to the official M14 accounting plan) - not entertainment expenses. Any

automated comparison between budgets that does not take this difference into account would be comparing apples to oranges.

2.1.9 G5.9 Contradicting data sources

Another problem journalists are facing when trying to access budget data are contradicting data sources. Two data sources seem to be describing the same thing but have different numbers and totals. In such cases, it is a hard job to pin down why they are different and which one is the correct one to use.

The American data journalist gives an example from his experience:

> *Sometimes you get several different sources what supposedly is the government spending and you get different answers and you have to look into the definitions how they are defining government spending, what are they including. [...] I will use an US example: the Congressional Budget office might have some tables and the Bureau of Economic Analysis might have some tables and they don't match. And suddenly it becomes a big job to find out why they don't match: Maybe one is using fiscal year, the other one is using calendar year. Or one has a different definition of social services as the other. It would be convenient if someone would be doing this for us. At least providing a widely agreed upon standard for how this things should be accounted for.*

2.2 Understanding budget data

2.2.1 G5.10 Understanding basic terms and context

Workshops with journalists and journalism students have shown that basic notions of accounting (the meaning of income, expenditure, asset and liability, for instance, or the meaning of the planned and executed budget) were far from being understood properly.

Similarly, the legal and political context (what administration is responsible for what) of budgets is not known by most of the non-specialist journalists encountered within this work package.

2.2.2 G5.11 In-kind spending and gifts

Local authorities can support third parties by offering them access to public resources for free. Football clubs, for instance, can rent stadiums for free or for a very low fee (TFT, Strasbourg, France; TFT, Basel, Switzerland).

To find the cost to taxpayers, journalists use estimations of the actual cost (the cost at which a stadium is rented to other companies, for instance) to derive the actual cost for the public authority. However, there is no financial transaction other than the rent (counted as income for the public body) and the actual subsidy remains intangible. To further complicate matters, some public bodies (TFT, Nice, France) do account for such in-kind subsidy precisely in their books, some other do not.

Similar issues arise when estimating the value of in-kind participations to a project. In Neuchâtel (TFT, Switzerland), the city “exchanged” a parcel valued at CHF 36m for a stadium of the same value, resulting in a neutral operation for the city. However, the rationale for setting the parcel’s value was not found. In Nice (TFT, France), the city ceded plots of land for EUR 1, a value deemed appropriate by the state authorities.

2.2.3 G5.12 Off-balance sheet operations

Public administrations can take on liabilities, such as guarantees offered to third parties (TFT, Hannover, Germany) or, more often, debt in the form of public-private

partnerships which are financed by private actors and rented by public authorities (TFT, Nice, France).

Such commitments would amount to non-trivial sums, were they contracted from private insurance companies or banks. However, there is no way for journalists to estimate them at market prices because there is no open market for insurances of such amounts.

2.2.4 G5.13 Understanding of accounting legislation

Understanding budget data requires a fine knowledge of specific accounting rules. Several examples were found in the case studies where, to get to the story, knowledge of accounting legislation was needed:

- Amortization: Investment expenditures can strongly skew budgets depending on amortization law. Investments can be amortized over any number of years, from a few to a few dozens. It is not rare that public administrations fail to follow the legal procedures for amortization, either purposefully (TFT, Strasbourg, France) to change their balance sheet one way or the other, or by mistake.
- Emphyteutic lease[7]: The city can lease a piece of real estate to an organization for a period of time (usually 99 years), this is called an emphyteutic lease. A hidden subsidy can take the form of the resiliation by the administration of an emphyteutic lease in exchange for hard cash, booked under exceptional expenditures. The city then rents the same terrain to the organization for a small fee. It appears in the books as a net gain in operating income when it is in fact a huge loss for the city (TFT, Strasbourg, France).

2.2.5 G5.14 Analytical accounting

Analytical accounting is a way to keep track of expenses and income according to a mission and not according to the legal accounting plan (as regular accounting does). Analytical accounting is done by most private companies to find the cost of a specific programme (e.g the cost of developing a new product). Public authorities also engage in analytical accounting for the same reasons. Journalists assessing the cost of a policy also do analytical accounting. However, such an endeavor is extremely costly to achieve based on currently available budget data.

The European Commission, for instance, listed in 2013 the expenses of all member states in Libya since the fall of Gaddafi. However, this list was to be found in a restricted document which was leaked to the public (TMF).

Moreover, it can be impossible for journalists to engage in analytical accounting based on public data. A single public authority can be in charge of several missions and not split its accounting between missions. An example from TMF (Baden Wüttemberg, Germany) showed that the administration kept tabs on people being forcibly removed in a single file. However, they did not fully specify if people were deported because of immigration (deportations) or criminal (extraditions) law.

2.2.6 G5.15 Difference in definitions

Interviewees and case studies showed that the semantics used by the administration could be far removed from the words used by journalists to describe the same issue. One case study, for instance (TMF, all countries) showed that the administration

[7] *Erbbaurecht* in German, *bail emphytéotique* in French, which is similar to *droit de superficie* (more used in Switzerland).

could use words such as “reception center” to describe a detention camp[8], for instance.

A French journalist faced the same issue when assessing the amount of government spending on environment-friendly programs. The government counted the tax rebates on propane and butane as “environment-friendly” even though the latest legislation clearly mentioned that fossil fuels could not be described this way. Similarly, the French government proposed that funds earmarked for cultural support be used to buy surveillance equipment around cultural buildings instead[9].

2.3 Training on Open Budget Data for journalists

The development from computer-assisted reporting to data journalism makes clear that numeracy (the ability to understand numbers and arithmetics), coding, data visualizations and a deep understanding of data is an essential part of journalism. Simultaneously, journalists need specific expertise on certain topics. Therefore, journalism schools and trainings are embracing in-depth training. The Master’s degree of the City University of London in Financial Journalism, for example, prepares journalists to cover business and the economy.

In this chapter we first describe different possibilities of data journalism training and financial journalism training. Our research shows a gap in training that combines both fields. In the sphere of journalism training there is a lack of training opportunities combining data journalism skills and expertise in budget data.

2.3.1 Existing Resource: Data Journalism Handbook

Published in March 2012, the Data Journalism Handbook is still probably the most used resource to learn about data journalism. The online handbook is free and available in English, Spanish, French and many other languages. The handbook serves as a manual, providing step-by-step training for beginners and more advanced users. It combines contributions from journalists, programmers, open access specialists and many other experts in the field and helps professional data journalists with a variety of backgrounds like programming, graphic design or journalism.

2.3.2 Existing Curriculum: School of Data

School of Data is a global network of educational programs offering training in data skills to journalists, civil society organisations and citizens in general. Born out of the open data movement, the founders of School of Data perceived a literacy gap in working with data. Therefore School of Data focuses on working with journalists as “multiplicators”. Journalists, on the other hand, benefit from this partnership and training as it advances their employability.

School of Data focuses on teaching beginner-level to medium data skills to practitioners who can easily put the new skills to use through the production of news stories. More advanced data training and specialized fields like advanced statistics, data science and programming are not the core of the School of Data curriculum.

Data training offered by School of Data varies from country to country. School of Data is pursuing a governance network structure, which leads to data training tailored towards specific needs in different countries and areas. For example, School of Data

[8] The CDA of Lampedusa is one example of such camp, where asylum seekers are refused the right to go outside, being in effect detained.

[9] See *Après les attentats de Paris, premières mesures de l’Etat pour les lieux de culture*, in La Gazette des Communes, 23 November 2015.

Brazil is working with big newsrooms to train their employees. Currently, it is teaching the journalists of the economics section of the biggest newspapers how to work with spreadsheets.

Methods from School of Data include in-person training sessions over a couple of days, online tutorials, data expeditions and fellowship programs. We highlight here their most successful model: the Data Expedition.

Data Expeditions

A typical data expedition starts with a thematic introduction to the issue. Participants indicate their skill level on the following terrains: design, storytelling, finding data, data-analytics, coding/programming (note that programming is only one of the skills).

Teams consisting of participants with mixed skill sets will together define the specific question or problem that they want to answer. Together they will have to identify and find the needed data and work their way through the data to come up with an answer or solutions.

Solutions and answers can come in different shapes or forms. A story or a sketch of the analysis is just as valuable as a fully functioning interactive map. Key is that the answer are based on the data that they have found and analyzed.

Data expeditions centralize the mutual learning over the outcomes. They focus on the data and the analysis and especially how to get from a problem to the solution independently.

The format was developed because the school of data teams missed this problem-solving aspect in then existing tutorials and workshop formats. To create a playful and enjoyable experience, the concept of data expedition was inspired by the Dungeons & Dragons role-playing game.

It is especially adequate for data journalism training, because journalists start a story with a hunch, question, or hypothesis. A data expedition teaches them how to go from there.

Training followed by WP participants at the at Ottowa, May 2015

Although intense in-person training is considered the most successful way of training, offering courses and resources online also serves another purpose.

> *The goal of online courses is more about creating communities around the subject than anything else. A lot of people give up on online courses; usually only ten percent get to the end - sometimes even less than that. But it is not about people finishing the course. The goal is to build a community who is interested in this topic. A community that is sharing experiences. The best approach is to use the online course to inspire people to build up the knowledge and a learning community.*
>
> – A Brasilian data journalism trainer

School of Data tries to keep the entry barriers for participants as low as possible and usually doesn't charge a fee. To keep the motivation up for students and to guarantee a high rate of attendance and completion in its classes, School of Data came up with a sophisticated scheme: Potential students need to apply and upon selection they need to sign a letter of commitment. Although it is legally not binding, this symbolic ritual leads to very positive results.

Specifically School of Data is offering an online course module in Budgets and Spending Data and also in Data Journalism.

2.3.3 Existing Curriculum: Budget Training at the City University of London Department of Journalism

Use case: A professor explains how he teaches how to cover the budget at a British university.

Learning how to cover the UK budget
At City University the students are spending a whole term on how to cover the budget. The term starts out with a set of dedicated classes and leads to a student group project covering the UK budget live on November 25. The goal is to imitate the intensity of a live coverage situation in a newsroom.

First, the students are attending five 3-hour workshops on numeracy. "I can't assume students know anything as they are coming from different backgrounds," said Steve Schifferes, Professor for Financial Journalism and former BBC economics editor. Topics covered include: numbers are your friends, a sense of size and scale, why you shouldn't be scared of numbers, how do you measure GDP and other indicators, world economy, basic statistics, how to lie with statistics.

In the fourth week the students start working on a group project that eventually leads to the live coverage of the UK budget which is announced on November 25, 12:30pm.

Each student starts out researching a specific topic like the economic forecast, the job situation, tax credits, health spending or the housing market. Then each student has to pitch their story idea in 2 minutes. The stories will be commissioned and the students have to work on the in the week ahead of the budget. The feature stories on those topics will be published before the budget announcement. With this background knowledge they will know what to look out for in the budget statement.

On budget day, the students will be assigned different roles in the live coverage: writing the lead article, writing the background article, live blogging, copy editing, social media, graphics and so forth. The students come in in the morning and start the live coverage. As soon as the budget is published at 12:30, they are live on it, publishing everything online.

When the day is over, we regroup and compare our coverage to what the professional newsrooms like BBC, the Financial Times, Reuters and others did.

Overall the goal is to learn about the budget by reporting on the budget.

2.3.4 Other existing online courses and resources

Beside the the Data Journalism Handbook and the resources from School of Data, the other major online resource in data training are Massive Open Online Courses (MOOCs).

Knight Center for the Americas at the University of Texas in Austin offers free course where everyone can participate from around the world with an Internet connection. The center is offering a variety of changing courses like *Introduction to Infographics and Data Visualization* or *Intermediate D3 for Data Visualizations*.

Further, the European Journalism Center (EJC) is offering a basic online video course in data journalism introduction called *Doing Stories With Data.*

2.3.5 Gaps in Training on Open Budget Data for Journalists

> *I think part of the problem is that outside of the data specialists who specialize in finding these resources there is still in this age a lot of reports who don't think as data as a source. There are more concerned with getting human source. Or if you are lucky, knowing a little about documents. I think they do not routinely monitor what data source are available and whenever the data is update check how the data has changed. I think this is overall an issue with journalism education.*
>
> – An American data journalist

Agreeing with this comment, we perceive a gap that journalism education is not emphasizing enough on data literacy and data visualization skills. First, in primary journalism education like universities and journalism schools, data journalism is at maximum a minor part of the curriculum. Second, there are opportunities in postgraduate and advanced training but they are often costly or time intensive. Further practicing journalists often do not realize how they work could benefit from data skills. Therefore there is a need for more data journalism training.

> *I come from a background where I work with National Computer Assisted Reporting which does all kind of data journalism training. And generally why this training is more effective than just say, buying a book for Excel or Access is because the exercises we teach are specifically tied to story scenarios. I think that in something like this and if you were trying to do training and you hear something very specific you can do with this table, you know, people pick it up much more quickly.*
>
> – An American data journalist

While data skills are a scarce resource in journalism, specific budget data skills are nearly non existing. Those two skills need to be taught together. With an increasing number of datasets concerning budget data available, the need for budget data training is increasing. Currently, there are opportunities to learn about data journalism and to learn about budget and spending data. But there is no offering teaching those two fields combined in one curriculum: doing budget data journalism. And even more so, there are no resources teaching how to use budget data to produce real world stories in journalism. Concluding, there is a gap in guiding journalists in budget data literacy and how to generate stories out of budget data.

Specifically, there is a need in more in-person trainings and tutorials as most resources are web based with a limited impact. As software, online tools and databases are changing frequently, written online resources are quickly outdated. In-person trainings are therefore more flexible in adapting to new developments and trends in the field of data journalism and budget data.

2.3.6 Learnings from the first on-site Workshop

In the realm of this gap analysis, we have conducted the first pilot workshop on-site at the Alter-EU conference on 20 and 21 June. We used the data from the Migrant Files project: the file with the dead counts on the mediterranean and the datafiles on the cost of migration. The audience consisted of 6 NGO-workers from different European countries. The goal of the training was to produce maps based on the different data files.

We intended to produce some very simple maps, such as point-maps (as the one produced by the migrant-files project itself) or thematic maps with the financial data, showing in which countries the biggest receiver of research money were located. To produce these maps we needed to address the following skills: understanding and

cleaning data, working with spreadsheets, geocoding, working with shapefiles, points and polygons.

However, the skill-level in the group was such, that we adjusted our goals after the first day to understanding and cleaning data, and working with spreadsheets. This taught us the important lesson that for us common operations such as converting csv-files, cannot be assumed common-knowledge.

In the first day we focused on the explanation of the topics, the methodology and how to read the datasets. This taught us the importance of documentation and methodology. Especially participants who had limited experience with data and statistics needed coaching in what you can conclude from the available data, and how to work with limitations. From this we draw that documentation of the data, and explanation of methodology and basic statistics to understand that methodology is key.

In the second session we asked: what would you like to know from these datasets. The first surprise for us was that they focused mostly on the dataset with the numbers of deaths in the Mediterranean sea. We had hoped to work primarily with the spending-data files. On the second day, interest rose when we provided more context and made the data concrete. For example, we showed them that money that is earmarked for migration flows back to research conducted by major companies such as Airbus. It taught us that spending data needs additional context so that it its implications can be understood.

The second surprise was that they asked questions that could not be answered with the dataset that we had available there - although we had extensively explained the limitations of the data. The participants wanted to be able to contextualize the data. For example, they wanted to know the number of deaths in relation to the overall refugee streams. The most important learning was to bring more contextual data.

The participants learned that not everything can be answered with hard data and some things must be estimated. For example, we cannot know how many refugees and migrants set off to reach Europe. We only know for sure who arrives. Although we had thoroughly explained the limitations in the data, the questions show that what people actually want to know often goes beyond the data. It remains important to teach how to get the data.

The workshop resulted eventually in basic graphs that were made in spreadsheet programs. We did not teach skills to the extent that we had expected, instead we learned the importance of explaining skills as understanding methodologies, basic statistics, asking the right questions etc. Second, spending data must be taught in its context. You need to show why it is interesting by linking to major company names that they know, or by highlighting discrepancies that they have to watch out for.

2.3.7 Summary Tables of Training Forms and their Pro's and Con's

In section 2.3 we have taken stock of the existing resource and training forms that are currently used for data-journalism and financial journalism. In the following tables we have summarized the different training forms to make an informed decision in the design of the workshops, tutorials and courses.

	Pros	Cons
Online - general	• Wide reach • Content can be obtained at any time from any place • Resources can be used exactly when needed	• Little opportunity for interactive learning • Difficult to obtain feedback • High drop-out rate
Guides & Manuals *Explainers to specific tools that guide you through their use.* *This can be in different forms: written, as a demonstrator, a demonstration website, or a step-by-step guide*	• Relatively easy to put together • It tells users exactly what to do with one tool	• Very limited use: only applicable to one tool • Most people will not take the time to go through them and want to start hands-on • Not interactive • High tool volatility let these resources go out of date easily
Tutorials *Short clips (+/- 10 minutes) that give an introduction to a certain process, or skills or one type of expertise (for example mapping explaining not just one software program, but also the principles of coordinate systems etc.)*	• Users can access them when they need them • The short duration makes it attractive for users • The explanations go beyond just "one" tool and show a process so that people can pick up a skill	• Not interactive • Limited, only addresses a specific skill • Hands on, will show how to start up a process • Usually still bound to one tool because the same functions differ that much between programs, that it is hard to write a comprehensive tutorial
Webinars *Webinars are online seminars that are usually recorded live. They can be on a wide range of topics and usually last 30 - 60 minutes. Example: IBP Seminar Series*	• Interactive • Used for conveying context and text-book knowledge, not hands-on skills • Can be watched afterwards	• Not hands on • Requires lengthy time-investment • Expensive to produce
MOOCs *Massive Open Online Courses. Online courses, combination of seminars, tutorials, assignments. They do not teach on*	• Allows one to teach a comprehensive course • Skills and context can be jointly taught	• Expensive to produce • High drop-out rate (approximately only 10% of the students completes the course

Project funded by the European Union's Horizon 2020 Research and Innovation Programme (2014 – 2020)

skill or one topic, but can encompass a course. Best example: data-science from Coursera.	• One can go in depth • Assignments let the participants practice • Knowledge intensive and hands-on	• Requires high commitment and time-investment from users (8-10 hrs p/w)

Table 1 - Pros and cons of online trainings

	Pros	Cons
Offline - general	• Direct feedback • Hands-on • Mutual learning and interaction between participants	• Participants or trainers need all to be at one venue • Language boundaries • High overhead for organization
Workshop *A workshop is a session where the focus is on training*	• Focus on knowledge transfer -> one skill can be taught extensively • Works well in groups with similar skill levels and limited time • Can be tailored for a specific group's needs	• The limited time only gives participants an appetizer of the possibilities • It is difficult to teach on skill for mixed groups
Data Expedition *One- or two-day events where diverse groups of participants will solve problems in teams working. The process is more important than the outcome.*	• Focus on mutual learning experience of participants • Participants learn from each other • Unexpected learning outcomes • Let's participants explore the entire process from questions, to data, analysis and presentation • Skills can be used independently afterwards • Outcomes can be diverse • Flexible format	• The theme-structured events do not always match daily practice • Little control over what participants learn

Hackathon or Datadive *Groups of programmers, data-scientists and experts will work on one program for a day or weekend. The goal is to produce prototypes*	• Experts will work on a specific problem, solution oriented • Room for creative, out of the box solutions • Concrete results that can be taken forward	• It is usually done 'for' an organization instead of 'with' the organizations • Focus is not primarily learning
Summer School/Intensive Course *Courses of several days that can combine workshops, seminars, data-dives and data-expeditions*	• Allows one to teach a comprehensive course or work on a project • Skills and context can be taught in depth • Knowledge intensive and hands-on • Very motivated participants • One can use a mix of teaching forms	• Time-intensive: many professionals cannot invest multiple days
Weekly Course	• Allows one to teach a comprehensive course or work on a project • Skills and context can be taught in depth • Assignments can deepen the independent application of learned skills • One can use a mix of teaching forms • Skills can be directly applied in the daily practice of participants • Professionals can more easily find weekly time in their schedules than consecutive days	• High drop-out rates • Time-intensive • Weekly courses require constant repetition, progress may be slow

Table 2 - Pros and cons of offline trainings

3 Conclusion and recommendations

3.1 The methodology

The alliance of case studies, workshops and interviews enabled this WP to collect information on 14 gaps that prevent journalists from working effectively with budget data - in a little over 5 months (work ran from June to November, with the interruption of the summer break). The methodology chosen is very efficient.

The WP will continue to collect feedback and test the solutions found to resolve the gaps through researching the case studies during workshops and in between them. The WP might launch a new case study, when the current ones to stop yielding insights regarding the way journalists work with budget data.

3.2 Improving the training on budgets data

Many of the gaps identified in this report find their origins in a lack of skills for data-journalism or lack of knowledge around public budget and expenditure processes. To address these gaps in a comprehensive manner, we will develop modules and workshop models that combine financial journalism and data journalism training.

As the tables above show, the different training forms in on- and offline format both have their pros and cons, which is why we will do a mixture of online and offline training. However, the expenses for good online training are such, that we will initially focus on in-person training to develop, test and tailor the course material. In addition, in-person training allow us to collect immediate feedback.

3.2.1 Design

We use the principle's from user-centred design to build up a curriculum for data- and financial journalism. User centered design focuses on the involvement of the users in all stages of the product development. We incorporate these principles by using the gaps from the case-studies that are simultaneously developed in this work package. Second, the curriculum will be developed in a series of pilots where we test out different course forms and sharpen our materials. Finally, the feedback phase allows us to revisit the strategy, and adjust the course material. The goal is to create a curriculum that can be used during and beyond the OpendBugets.eu project.

Research and Outreach Phase

The interviewees related that the success of the training will not only lie with the quality and content of the material, but also with finding motivated participants, who have the backing of their editors (if they are employed). During this phase, the research will focus on the specific needs of the stakeholder group and prioritize skills to shape the curriculum.

Pilot Phase: Sprints for workshops, one-day expeditions, and one multiple-day intensive course

In the second phase we develop the first concrete course material. We work together with media-organizations, journalism schools, and conferences to conduct on-side training. We will first develop short workshops and organize the first data-expeditions. The goal is to work towards a three-day intensive training that can be held in par with journalism conferences.

Project funded by the European Union's Horizon 2020 Research and Innovation Programme (2014 – 2020)

Feedback and Reflection phase:

In the third phase we reflect on the learnings from the pilot phase and revisit the course material. This phase may result in consolidation of the existing course material and may bring new gaps to light that will result in another testing-phase.

3.3 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).	Start of the resolution action.
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).	Start of the resolution action.
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).	Start of the resolution action.
G5.4 Timeliness	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Start of the resolution action.
G5.5 Completeness and level of details	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Start of the resolution action.
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.	Start of the resolution action.
G5.7 Identification of co-financing	Increase awareness for transparency among public bodies (collaboration with WP6).	Start of the resolution action.
G5.8 Lack of standardization	Increase awareness of the issue among public administration bodies (collaboration with WP6).	Start of the resolution action.
G5.9 Contradicting data sources	Tutorials for journalists and public administration officials.	Start of the resolution action.
G5.10 Understanding basic terms and context	Tutorials for journalists.	Start of the resolution action.
G5.11 In-kind spending and gifts	Addition of in-kind spendings to the OB ontology when possible (collaboration with WP1). Tutorials for journalists.	Start of the resolution action.
G5.12 Off-balance sheet operations	Addition of off-balance sheet operations to the OB ontology (collaboration with WP1). Tutorials for journalists.	Start of the resolution action.
G5.13 Understanding of accounting legislation	Tutorials for journalists.	Start of the resolution action.
G5.14 Analytical accounting	Tutorials for journalists.	Start of the resolution action.

G5.15 Difference in definitions	Tutorials for journalists.	Start of the resolution action.