# European Structural Funds: A Data Quality Index.

***Michael Peters***
***Anna Alberts***
***Bela Seeger***
*Open Knowledge Foundation Germany*

The project has received funding from the European Union's **H2020 EU research and innovation programme** under grant agreement **No 645833**.

This report was written and researched by **Michael Peters, Anna Alberts, and Bela Seeger** of **Open Knowledge Foundation Germany** on behalf of **OpenBudgets.eu**, a Horizon 2020 research project executed by **Fraunhofer IAIS**, **Open Knowledge Greece**, **Fundación Ciudadana Civio, Transparency International EU Office**, **Open Knowledge Foundation Germany**, **Vysoká škola ekonomická v Praze**, **Journalism++**, and **Universität Bonn.** It has been made possible through additional funding by **ADESSIUM Foundation**.

OpenBudgets.eu is a Horizon 2020 funded project that aims to provide a generic framework and concrete tools for supporting financial transparency, thus enhancing accountability within public administrations and reducing the possibility of corruption.

This report evaluates the data collected for subsidystories.eu - a project funded by Adessium and OpenBudgets.eu. Subsidystories.eu collected all available beneficiary lists for European Regional Development Funds, European Social Funds, and Cohesion Funds for the funding periods 2007 - 2013 and 2014 - 2020 across Europe.

Find out more at http://openbudgets.eu/about/ and http://subsidystories.eu.

# CONTENTS


## EXECUTIVE SUMMARY

SUMMARY OF RECOMMENDATIONS

## INTRODUCTION

## I. POLICY BACKGROUND

## II. OBTAINING THE DATA

2.1 EVALUATION OF DATA PORTALS

2.2 RANKING PRACTICAL USABILITY

## III. DATA QUALITY

3.1 DATA FORMATS

3.2 RANKING DATA QUALITY

3.3 OVERALL RANKING

## IV. COMPARING DATA

4.1 AMOUNTS

4.2 EU VARIABLES

4.3 VISUALIZATION

## CONCLUSION

# EXECUTIVE SUMMARY

This report discusses the quality of the **EU member states' beneficiary data** released for the **European Structural and Investment Funds** for the funding periods of 2007-2013 and 2014-2020. Special focus is laid upon the accessibility of the data via the managing authorities websites and the quality and format of this data.

EU Regulation No 1303/2013 from December 2013 requires the member states to create a **single website** providing all viable information on their **operational programmes** and publishing their **beneficiary data in a machine-readable format**. For their previous project SubsidyStories, funded jointly by Adessium and OpenBudgest.eu, Open Knowledge Germany and Open Knowledge International collected all data for the 2007-2013 and 2014-2020 funding periods, which set the foundation for the analysis in this report.

***"while SubsidyStories.eu scraped and cleaned this data to make it accessible for everyone, it should have never been necessary, had the member states complied with the regulations."***

All EU member states' ESIF websites were analyzed and evaluated against the governing EU regulation with special attention towards usability, data access and their availability in English. We have concluded that Lithuania, Poland and Bulgaria had the most useful websites, with Poland on top of the ranking due to their clever use of illustrations. Overall only 16 of the 28 member states provide English translations to their websites, which makes access for other EU citizens difficult. Machine readability of data formats has improved substantially in the 2014-2020 period, with less and less PDFs being published. However, member states are still far from completely adhering to the EU regulation with only 23 of 28 countries having released their data as of February 2017. Furthermore, six member states still used close data formats such as PDF or specifically designed webapps, which do not allow for easy data extraction or comparative analysis.

While subsidystories.eu scraped and cleaned this data to make it accessible for everyone, it should have never been necessary, had the member states complied with the regulations. In short, the data quality has improved in the funding period 2014-2020, as compared to the 2007-2013 funding period, but much remains to be done.

# SUMMARY OF RECOMMENDATIONS

1. **Require member states to make websites available in English**
2. **Make CSV or JSON the mandatory format for beneficiary data**
3. **Include information on the legal form of the beneficiary**
4. **Require standardised date-notation**
5. **Provide standardised way to make non Euro amounts comparable**
6. **Provide the following amounts: applied, allocated, and paid out.**
7. **Provide project funding broken down by EU Amount, Member State Amount, Third Party Amount, and a Total Amount**
8. **Provide information on the following dates and milestones in the project: start, finish, payment date and duration**
9. **Provide sufficient information to link the beneficiary lists to the programmes by CCI codes**
10. **Provide sufficient geographical information for both beneficiary and project location**

# INTRODUCTION

This report was written for the EU financed project OpenBudgets.eu. The data that this report relies upon is based on earlier work by the Open Knowledge Foundation Germany and Open Knowledge International in their "Subsidystories.eu" project, jointly financed by ADESSIUM Foundation and OpenBudgets.eu.

In this project the ERDF, ESF and CF data for both the 2007-2013 and 2014-2020 period were collected for all EU member states. Data was mapped and visualized with the Open Fiscal Data package also used in the OpenBudgets.eu project and is open and available at www.subsidystories.eu.

The report is divided into four main parts: firstly some policy background will be provided, followed by a section on how the data was obtained including a ranking on the member states' data portals. Thereafter, the data quality will be evaluated and ranked according to the EU's criteria. Lastly, opportunities for visualizing and analyzing the EU's spending data will be discussed.

# 1. POLICY BACKGROUND

To give some context to what the European Structural Investment Funds are and how they work, the EU's investment policy will be discussed. The EU Commission laid out their Horizon 2020 strategy for generating smart, sustainable and inclusive growth in the EU. In order to achieve these goals, the EU manages the European Structural Investment Funds, which are its main investment policy tools. To assure that the funds are used to achieve the EU's goals, detailed investment priorities and thematic objectives are defined, which function as guidelines for the use of the funds. The European framework constitutes funding periods of seven years with the last period ranging from 2007-2013 and the current period lasting from 2014 until 2020.

Institutionally, the member states and the European Commission (through its directorates general) negotiate a Partnership Agreement within the benchmarks that are set by the regulations for the structural and cohesion funds. Partnership agreements are contracts governing the funding process between the European Commission and the member states. Thereafter, the operational programme (OP) have to be submitted based on how applicants are planning to achieve the Commission's goals by funding local projects. The applicants for these operational programmes are the member states' regions as defined by the NUTS classification (Nomenclatura of territorial units for statistics). Within the regions a management authority has to be declared such as ministries of finance or regional administrations. While application is always handled by the region, countries with a strong central state often administer the funds on a national level. This leads to spending data being released on a national level. For countries with a federal structure such as Germany, Spain and Austria, data is usually published on the regional level.

The member states have to give detailed descriptions on their goals and how they plan to achieve these with the respective ESIF funds. Goals have to be in line with the thematic objectives and investment priorities published by the European Commission. After submitting the OP, they are reviewed by the responsible directorate general (DG). If accepted, the management authorities receive the funds from the DG and use their own websites to advocate funding. Thereafter, individual project application starts. Our investigation on available datasets has already shown that some countries are rather slow on the application side, because they still have not published any data for the 2014-2020 period (Cyprus, Malta, Spain).

The European Structural Investment Funds (**ESIF**) cover five different instruments:

- European Regional and Development Fund (**ERDF**)
- European Social Fund (**ESF**)
- Cohesion Fund (**CF**)
- European Agricultural Fund for Rural Development (**EAFRD**)
- European Fisheries Fund (**EFF**)

With subsidystories.eu, we focused on three of these ESIF funds: The ERDF and Cohesion Fund managed by the Directorate General for Regional and Urban Policy and the ESF overseen by the Directorate General for Employment, Social Affairs & Inclusion. While the ERDF aims to strengthen economic and social cohesion in the European Union by correcting imbalances between its regions (here), the ESF is Europe's main instrument for supporting jobs, helping people get better jobs and ensuring fairer job opportunities for all EU citizens (link).

***„Our aim is to improve fiscal transparency in the European Union by fostering the access to its spending data and allowing for cross country comparison for the first time"***

While all member states can apply for ERDF/ESF funding, the Cohesion Fund only applies to member states whose Gross National Income (GNI) per inhabitant is less than 90 % of the EU average. For the current period this concerns: Bulgaria, Croatia, Cyprus, the Czech Republic, Estonia, Greece, Hungary, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovakia and Slovenia.

While the EU provides spending data on the aggregate (member state or regional) level, this project gathered all available data on which beneficiaries receive European funding and which projects are implemented. Our aim is to improve fiscal transparency in the European Union by fostering the access to its spending data and allowing for cross country comparison for the first time.

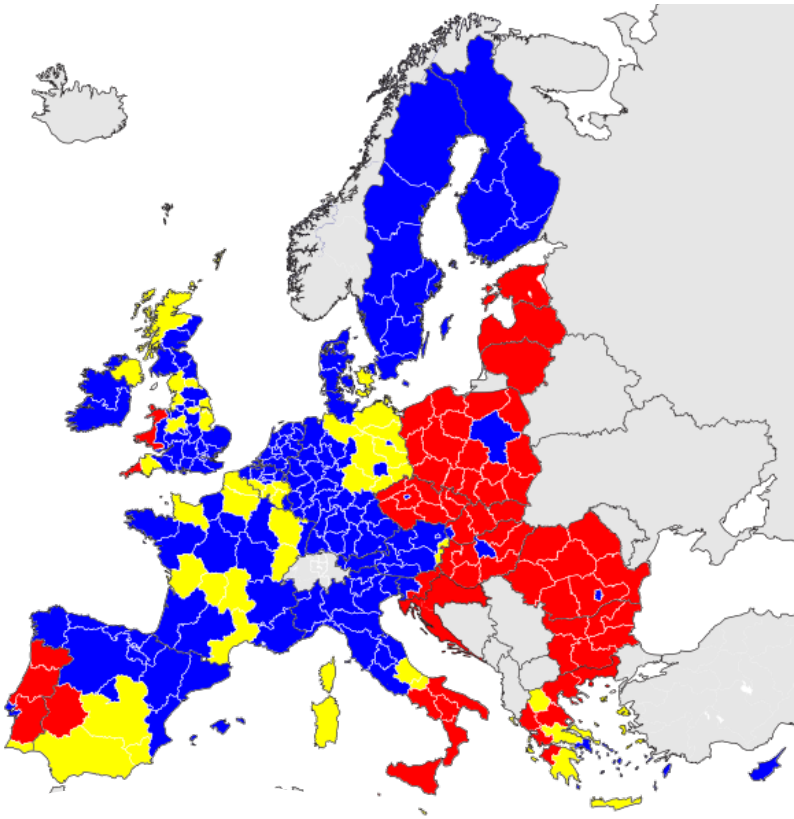

Figure 1.
Map of EU Regions
Classification of regions from 2014 to 2020:
- Less developed regions
- Transition regions
- More developed regions.
Source: Wikipedia.

# II. OBTAINING THE DATA

EU member states have been required to publish the data online since the 2007-2013 period. However, the 2007 regulation was still vague and lead to some member states publishing detailed datasets, while the majority only published basic information on beneficiary names, amounts and dates. The management authorities usually create a website regarding the European Structural Investment Funds (ESIF), where they offer information on funding opportunities for possible beneficiaries and list previous projects etc. In some cases, this means there is one website / online portal, where information on all funds (ERDF, ESF and CF if applicable) is provided such as France, Cyprus or Denmark. In countries with a decentralized state - like Germany, Austria and Belgium - regions function as management authorities and hence, publish the data on a regional website. For Germany's 16 regions this leads to 16 different websites, however, the websites are often separately distinguished by funds, meaning the actual number of websites for Germany is 27. You can find an overview on the country specific portals below:

**Table 1: Overview Data Portals**

| | |
|---|---|
| Austria | http://www.efre.gv.at/projekte/projektlandkarte/ |
| Belgium | http://www.vlaio.be |
| Bulgaria | http://umispublic.government.bg |
| Croatia | http://www.strukturnifondovi.hr |
| Cyprus | http://www.structuralfunds.org.cy |
| Czech Republic | http://www.dotaceeu.cz/cs/Informace-o-cerpani/Seznamy-prijemcu |
| Denmark | https://regionalt.erhvervsstyrelsen.dk/ |
| Estonia | http://www.struktuurifondid.ee/programming-2014-2020/ |
| Finland | https://www.eura2014.fi/rrtie-pa/?lang=en |
| France | http://www.europe-en-france.gouv.fr |
| Germany | http://www.esf.de/portal/DE/Startseite/inhalt.html |
| Greece | https://www.espa.gr/en/pages/default.aspx |
| Hungary | https://www.palyazat.gov.hu/ |
| Ireland | http://eustructuralfunds.gov.ie |
| Italy | http://www.opencoesione.gov.it |
| Latvia | http://www.esfondi.lv/es-fondu-projektu-mekletajs |
| Lithuania | http://www.esinvesticijos.lt |
| Luxembourg | http://www.fonds-europeens.public.lu |
| Malta | https://investinginyourfuture.gov.mt/projects?lang=mt |
| Netherlands[2] | https://www.europaomdehoek.nl |
| Poland | http://www.mapadotacji.gov.pl/en |
| Portugal | https://www.portugal2020.pt/Portal2020 |
| Romania[3] | http://www.inforegio.ro/ |
| Slovakia | https://www.itms2014.sk |
| Slovenia | http://www.eu-skladi.si |
| Spain | http://www.dgfc.sepg.minhafp.gob.es/sitios/dgfc/en-GB/Paginas/inicio.aspx |
| Sweden | http://projektbank.tillvaxtverket.se/projektbanken2020#page=eruf |
| UK - England | https://www.gov.uk/government/publications/ESIF-useful-resources |

1,3 - For The Netherlands, different files are available for the European Social Funds on national level, and for the ERDF on regional level in different formats and from different quality. For Chapter 3, we decided to only evaluate the data portals as indicated. However, in chapter 4 – 6, the data as eventually located was used in our evaluation.

The EU provides an overview on some of the websites in their own portal here. It is a good starting point, but not necessarily up to date. Online searches of "ERDF/ ESF + beneficiary + respective country/region" usually lead to the required portals. While some websites are available in English, others are not and require using website translation. Obtaining the data can therefore be quite troublesome.

## 2.1 EVALUATION OF DATA PORTALS

The following section focuses on evaluating all 28 European data portals and the ranking that we developed for this report. The ranking is based on criteria such as availability of the website in English, ease of use, functionality and how easily beneficiary data can be found. The regulation reads: "[...] giving examples of operations, by operational programme, on the single website or on the operational programme's website that is accessible through the single website portal; the examples should be in a widely spoken official language of the Union other than the official language or languages of the Member State concerned."

***„The first obstacle when confronted with a foreign countries data portal is usually the language, even though a "widely spoken official language" of the Union is required. "***

As discussed above we used the EU's own data portal as a starting point for our search, and if a specific website was not included, we searched for it. The first obstacle when confronted with a foreign countries data portal is usually the language, even though a "widely spoken official language" of the Union is required. Overall 12 out of the 28 countries do not provide any English assistance. This is problematic, because the websites have to be translated first, in order to allow for any further research. We used http://itools.com/tool/google-translate-web-page-translator for this task. It remains to be said, that even if websites offer translations, this does not guarantee their helpfulness. Often the translated pages just cover a small part of the original website and in some cases do not allow for finding the beneficiary data while in the English mode, such as the German and French portals.

Finding a coherent way of evaluating the country portals and the beneficiary data is difficult, due to their differences in conception. As discussed, countries with a strong federal state tend to distribute the ESIF funds on a regional level, leading to multiple and different portals. Some even have distinguished platforms for the ERDF and ESF. For the 2014-2020 period we looked closely at the ERDF data and respective portals, and noted if they included all or fund specific information. In case of countries that published the data regionally, we considered one regional dataset such as the Belgian region Flanders or the German region Berlin. However, it should be noted, that not all Belgian or German regions have published their data yet. In case there was no data available for the 2014-2020 period (Austria, Spain, Romania, and Cyprus), we still evaluated the webpages based on the 2007-2013 period.

# 2.2 RANKING PRACTICAL USABILITY

The scores depicted are a combination of a few simple questions that we wanted to be answered by the portal:

- **Was the website available in English?**
- **How easily could the portal be located by using Google search?**
- **How long did it take to find the beneficiary data?**
- **Could the data be downloaded directly or did it require scraping?**

These questions do have subjective nuances, e.g. finding the beneficiary data on the website can to an extent be fostered by luck of clicking on the correct subpage. However, this is influenced by the fact that the pages are available in English or follow a clear and intuitive structure. The subjectiveness of "ease of use" should be considered when viewing this ranking. Factors such as design or "look" of the website were neglected unless they specifically aided the access to beneficiary data. Furthermore, we are only considering the data format here and not the data quality, which will be evaluated by itself later on. Scores were awarded on a scale from 1-5 with one being the lowest and five the highest possible score. Countries that fulfilled all our criteria received a five, while minor issues led to a four, if no data could be found, websites could not be located or other major issues existed they received a one. Results are presented in table 2 (page 10).

Our benchmarks of practical websites are from Bulgaria, Lithuania and Poland. The portals can be easily found via Google and are all available in English. Beneficiary data can be located very quickly and then downloaded in a machine readable format. Additionally, the lithuanian and polish website offer useful illustrations (such as maps /charts) that give a general idea of the data. Overall the polish website http://www.mapadotacji.gov.pl/en is our winner (figure 2), because of its intuitive use and great illustrations, the easy data download and detailed English project descriptions. However, it is important that these illustrations remain an additional feature to the openly accessible machine readable datasets. Relying only on interactive maps that show where single projects are based and how much they costs was considered negatively. These webapps do not enable cross project comparisons and scraping the data to a machine readable format is very tedious.

***"Our benchmarks of practical websites are from Bulgaria, Lithuania and Poland. The portals can be easily found via Google and are all available in English"***

Bad practice examples come from Cyprus and Romania (figure 3, next page). While portals exists in these two countries, they are not translated to English and are therefore difficult to navigate. Even after using a website translation service they remain hard to use and there is no available data for the 2014-2020 period.

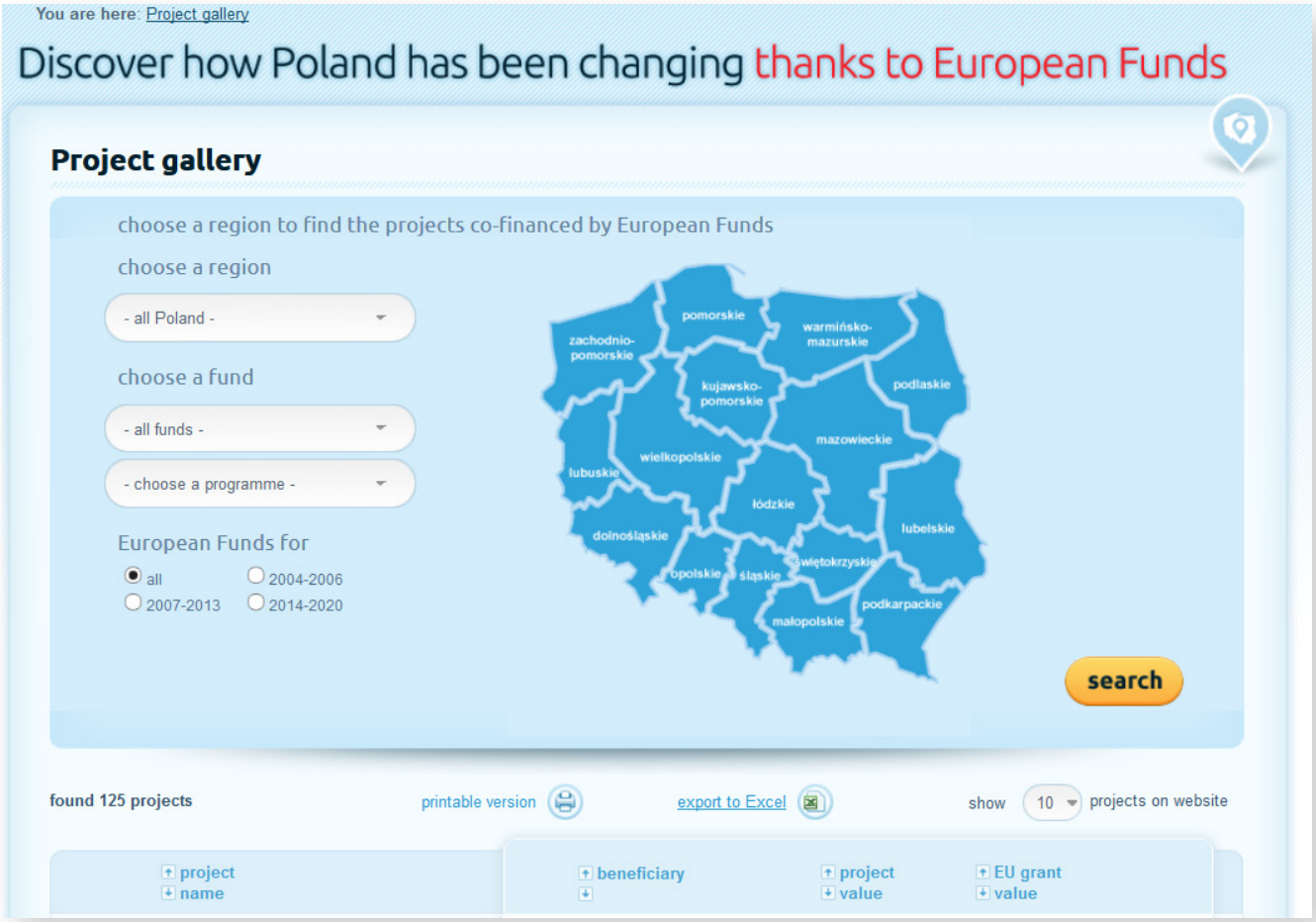


Figure 2.
**Polish Spending Data Webportal**
The winner of our comparison with intuitive use and great illustrations, easy data download and detailed English project descriptions.

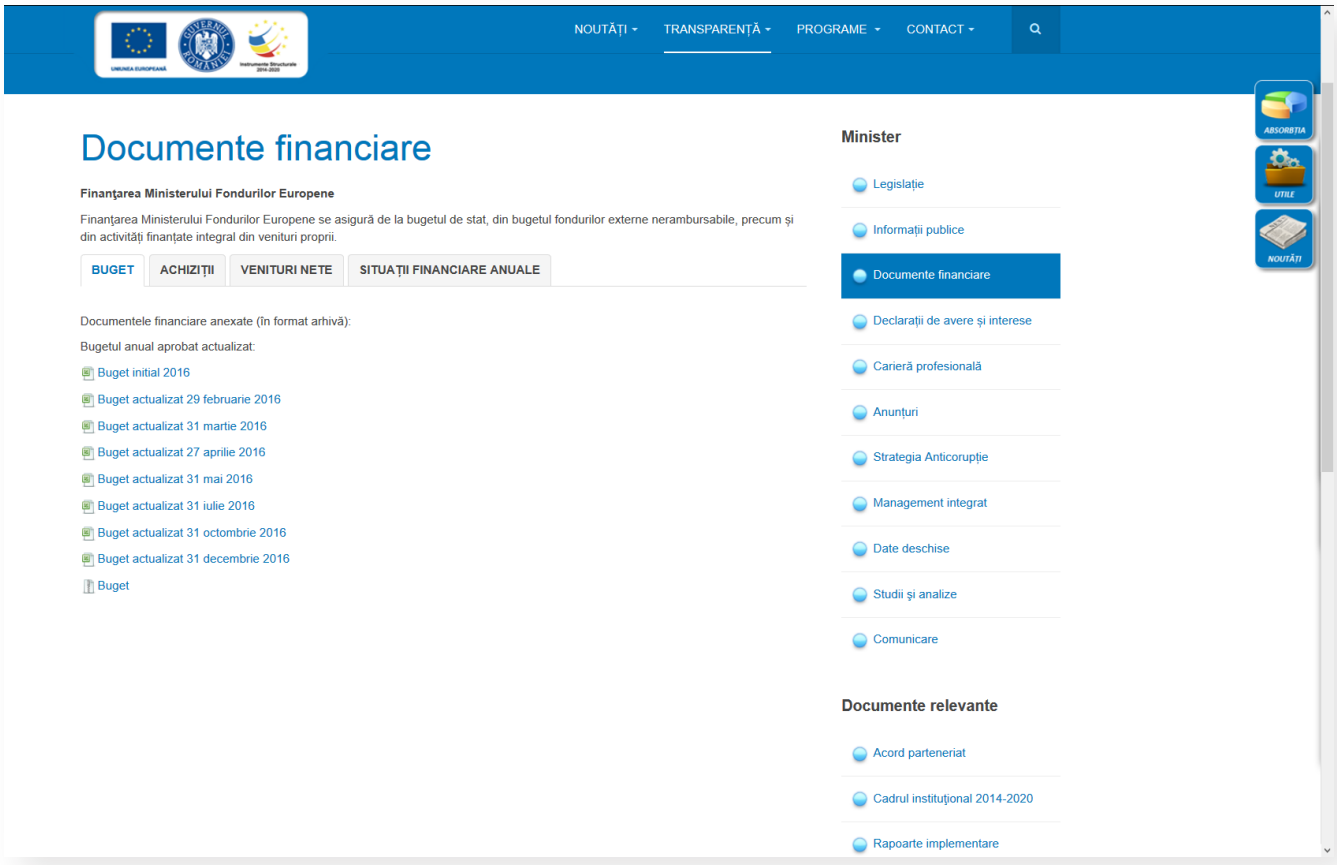


Figure 3.
**Romanian Spending Data Webportal**.
An example of a less well-designed portal.

**Table 2: Ranking Data Portals**

| COUNTRY | DATA FORMAT | ENGLISH TRANSLATION | SCORE |
|---|---|---|---|
| Bulgaria | XLS | yes | 5 |
| Lithuania | XLS | yes | 5 |
| Poland | CSV | yes | 5 |
| Slovenia | XLS | yes | 5 |
| Belgium Flanders | XLS | yes | 4 |
| Denmark | CSV | yes | 4 |
| Finland | CSV | yes | 4 |
| Greece | CSV | yes | 4 |
| Italy | XLS | no | 4 |
| Portugal | XLSX | no | 4 |
| Austria[2] | XLSX | yes | 3 |
| Croatia | XLS | no | 3 |
| Czech Republic | XLSX | yes | 3 |
| Estonia | webpage | yes | 3 |
| Hungary | webpage | yes | 3 |
| Latvia | XLS | no | 3 |
| Luxembourg | webpage | no | 3 |
| Slovakia | webpage | no | 3 |
| France | XLS | yes | 2 |
| Germany | XLSX | yes | 2 |
| Malta | PDF | yes | 2 |
| Romania | PDF | no | 2 |
| Sweden | webpage | no | 2 |
| UK - England | XLSX | yes | 2 |
| Cyprus | | no | 1 |
| Ireland | | yes | 1 |
| Netherlands[3] | XLSX | no | 1 |
| Spain | | yes | 1 |

2 - The Austrian Portal was updated and brought to our attention in April 2017.

3 - The Dutch portal is ranked with 1 point due to the fact that it only provides an excerpt of projects funded by the Netherlands.

# III. DATA QUALITY

In general, we can say that the data from the 2014-2020 funding period is substantially better and easier to access than the previous period. This is likely due to the fact that the new EU legislation "Regulation (EU) No 1303/2013 of the European Parliament and of the Council of 17 December 2013" mandated the form the data should be presented in. The data shall be uploaded in the aforementioned online portals in a machine readable format and at least include the variables: beneficiary name, project name, operation summary, start & end date, total eligible expenditure, union co-financing rate, operation postcode, name of category of intervention and date of last update. 2014 – 2020 data is not yet available for every member state, because some have simply not released it yet. Some countries like Italy have released information only on the level of operational programs, where no single beneficiaries are listed, because the projects are simply "not determined" yet. For similar reasons other countries have not released any data at all up to this point. We have collected all the data to our best knowledge and have inquired with the national / regional authorities if we could not find anything. Our research includes all data published until the end of January 2017.

## 3.1 DATA FORMATS

As discussed above, our research confronted us with many different formats in which the data was presented. This is despite the fact that the regulation for the 2014-2020 period clearly states that machine-readable formats shall be used (such as CSV). This is not the case for all countries as table 2 demonstrates. Out of the 22 countries that had uploaded their ERDF data, only 16 can be considered machine-readable formats. For this case we are counting XLS, XLSX and CSV as machine-readable, although only CSV truly is. However, XLS and XLSX can usually be converted to CSV rather easily.

**Table 3: Data Formats ERDF 2014-2020**

| FORMAT (BEFORE SCRAPING) | # | FORMAT (AFTER SCRAPING) | # |
|---|---|---|---|
| JSON | 0 | JSON | 6 |
| CSV | 4 | CSV | 16 |
| XLSX | 5 | XLSX | 0 |
| XLS | 13 | XLS | 0 |
| WEB | 6 | WEB | 0 |
| PDF | 37 | PDF | 0 |
| Total | 22 | Total | 22 |

However, getting the data out of the PDF format is a lot more tedious, since the data cannot be accessed directly. In order to extract data from a PDF the file has to be "scraped" – that is an automated way to obtain the information from the original file has to be found. This can be done by coding, if you are an experienced developer or with automated tools such as Tabula.

Another source of data are web portals such as the French 2007-2013 site (figure 4), which shows a map indicating which region/city/municipality received what amount of funding. While these maps are a good way of visualizing data, they hinder the use of the data. Comparing projects to one another is impossible, because single projects have to be selected. Furthermore, data cannot be aggregated and is difficult to retrieve, because it might be embedded in HTML. Our developer often spent several hours at a time coding to retrieve the underlying data. The two data formats that we are able to process in OpenSpending are Comma Separated Values (CSV) or JavaScript Object Notation (JSON), both completely machine readable. Therefore, all the other files had to be converted to that format.

To get an impression of the overall progress in data formats and a possible effect the newly introduced EU regulation might have had, table 4 is presented. It shows the distribution of data formats for the 2007-2013 and the 2014-2020 period. The number of datasets in machine readable formats has improved. This is most visible in the number of datasets presented in PDF (49 in 2007 vs. 33 in 2014) and XLSX (5 in 2007 vs. 31 in 2014). This is a positive development that we want to highlight, although many of the datasets do not comply with the self-prescribed EU standards.

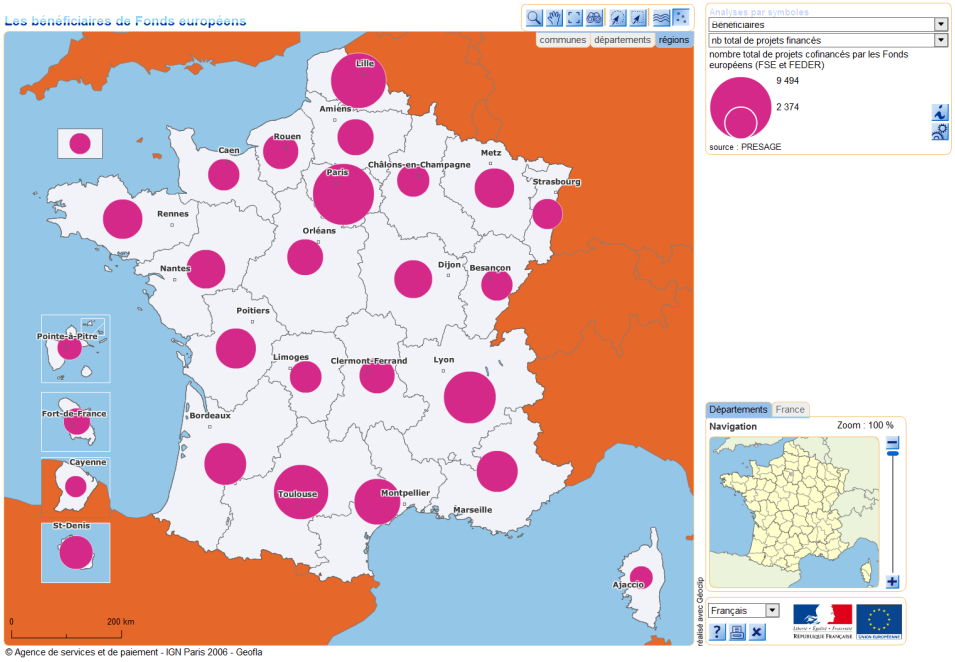


Figure 4.
**French Spending Data Web Application**.

**Table 4: Data Formats 2007 vs. 2014**

| FORMAT 2007 | # | FORMAT 2014 | # |
|---|---|---|---|
| JSON | 0 | JSON | 1 |
| CSV | 9 | CSV | 4 |
| XLSX | 5 | XLSX | 31 |
| XLS | 12 | XLS | 9 |
| WEB | 10 | WEB | 7 |
| PDF | 53 | PDF | 33 |
| Sum | 89 | Sum | 85 |

# 3.2 RANKING DATA QUALITY

In the table 5, all 28 EU countries are listed (by NUTS code) alongside the mandatory variables that the EU regulation requires. It can be seen that not all datasets are published yet (i.e. Cyprus and Ireland) and therefore could not be judged. Most countries that have published their data, comply with the new standards quite well.

**Table 5: [Ranking Data Quality]**

| NUTS CODE | FUNDS | BENEFICIARIES | PROJECT NAME | DATE LAST UPDATE | ELIGIBLE EXPENDITURE | START DATE | END DATE | CO-FINANCING RATE | POSTCODE | PROJECT SUMMARY | INTERVENTION CATEGORY | FORMAT | RATE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DE3 | • | • | • | • | • | • | • | • | • | • | • | xlsx | 11 |
| DK | • | • | • | • | • | • | • | • | • | • | • | csv | 11 |
| MT | • | • | • | • | • | • | • | • | • | • | • | pdf | 11 |
| UK | • | • | • | • | • | • | • | • | • | • | • | xlsx | 11 |
| AT | • | • | • | • | • | • | • | • | • | • | • | xslx | 11 |
| RO | • | • | • | | • | • | • | • | • | • | • | pdf | 10 |
| NL | • | • | • | • | • | • | • | • | • | • | | xls | 10 |
| EL | • | • | • | • | • | • | • | • | | • | • | csv | 10 |
| FR | • | • | • | • | • | • | • | • | • | | • | xls | 10 |
| HU | • | • | • | • | • | • | | • | • | • | • | web | 10 |
| LU | • | • | • | • | • | • | • | • | • | | • | web | 10 |
| SE | • | • | • | • | • | • | • | | • | • | • | web | 10 |
| SI | • | • | • | • | • | • | • | | • | • | • | xls | 10 |
| SK | • | • | • | • | • | • | • | | • | • | | web | 9 |
| HR | • | • | • | • | • | • | • | | | | | xls | 7 |
| FI | • | • | • | • | • | • | • | | | | | csv | 7 |
| BE2 | • | • | • | • | • | | | • | | | | xls | 6 |
| EE | • | • | • | • | | • | • | | | | | web | 6 |
| PT | • | • | • | • | • | | | | | | • | xlsx | 6 |
| BG | • | • | • | • | • | | | | | | | xls | 5 |
| IT | • | • | | • | • | | | | | | • | xls | 5 |
| LV | • | • | • | • | • | | | | | | | xls | 5 |
| LT | • | • | • | • | • | | | | | | | xls | 5 |
| PL | • | • | • | • | • | | | | | | | csv | 5 |
| CZ | | • | • | • | • | | | | | | | xlsx | 4 |
| CY | | | | | | | | | | | | | 0 |
| IR | | | | | | | | | | | | | 0 |
| ES | | | | | | | | | | | | | 0 |

# 3.3 OVERALL RANKING

To create an overall ranking for both the data and the websites, we combined the scores the countries obtained in both rankings. The website ranking offered scores from 1-5 judging website accessibility, English content and the data formats. The data quality ranking ranked the member states according to how well they fulfilled the established EU criteria. Member states had to publish 11 variables and are therefore ranked from 0-11.

In order to combine these two rankings for an overall index, we had to standardise the website ranking to make it comparable to the data quality ranking. We therefore multiplied the score from the website ranking by two, thereby ensuring that accessibility and data formats are equally considered. This lead to the overall ranking presented in table 6.

Slovenia wins the overall ranking, due to the openness of its data and the convincing website. Furthermore, it included 10 of the required 11 variables and therefore complies with the EU standards. Denmark comes in second, they provide outstanding data quality in machine readable format, but the website is hard to locate and access. Countries that have not publish their 2014-2020 beneficiary data yet, were severely punished in the rankings. That is the case for Austria, Cyprus, Ireland and Spain, their websites are not up to date, no new data is included and hence none of the required EU variables are provided.

**Table 6: [Overall Ranking]**

| # | Country | Overall Score | Score: Portal (2x) | Score: Data Quality |
|---|---|---|---|---|
| 1 | Slovenia | 20 | 10 | 10 |
| 2 | Denmark | 19 | 8 | 11 |
| 3 | Greece | 18 | 8 | 10 |
| 4 | Hungary | 16 | 6 | 10 |
| 4 | Luxembourg | 16 | 6 | 10 |
| 6 | Malta | 15 | 4 | 11 |
| 6 | Germany | 15 | 4 | 11 |
| 6 | United Kingdom | 15 | 4 | 11 |
| 9 | Slovakia | 15 | 6 | 9 |
| 9 | Finland | 15 | 8 | 7 |
| 9 | Bulgaria | 15 | 10 | 5 |
| 9 | Lithuania | 15 | 10 | 5 |
| 9 | Poland | 15 | 10 | 6 |
| 14 | France | 14 | 4 | 10 |
| 14 | Netherlands | 14 | 4 | 10 |
| 14 | Romania | 14 | 4 | 10 |
| 14 | Sweden | 14 | 4 | 10 |
| 14 | Belgium | 14 | 8 | 6 |
| 14 | Portugal | 14 | 8 | 6 |
| 20 | Austria | 13 | 2 | 11 |
| 20 | Croatia | 13 | 6 | 7 |
| 20 | Italy | 13 | 8 | 5 |
| 22 | Estonia | 12 | 6 | 6 |
| 23 | Latvia | 11 | 6 | 5 |
| 24 | Czech | 10 | 6 | 4 |
| 25 | Cyprus | 2 | 2 | 0 |
| 25 | Ireland | 2 | 2 | 0 |
| 25 | Spain | 2 | 2 | 0 |

# IV. COMPARING DATA

Negative outliers are Bulgaria, Czech Republic, Estonia, Latvia, Lithuania and Poland. These countries fulfil less than six of the required ten data fields. The Czech dataset offers the least amount of required data, but is at least available in an open format contrary to the Estonian data which had to be scraped from a webpage.

High quality datasets come from Denmark, Germany, France, Slovenia and the UK. These countries complied with all the standards, often adding further data fields. Furthermore, avll above stated files were available in an open format which makes them more accessible and easier to compare. The best dataset was offered by Denmark, because it complied with all categories and was published in CSV, making it even better than the English and German (Berlin) datasets that included all information but were only available in XLSX. A further advantage of the Danish dataset vs. the German data is that it comes in one dataset and not in different regional editions. This is a huge disadvantage of the German data, which is furthermore distinguished by ERDF and ESF, making data collection more tedious.

In order to create a large database including all data from all European countries for the ESIF funds, we had to find a common denominator for understanding fiscal data. We used a modified version of Open Spending's fiscal data model to map (unify) the data. The collected data confronted us with two major issues: format (discussed above) and content. We had to find a common denominator to enable comparing projects across different countries, guaranteeing that an amount in the Italian dataset can actually be compared to an amount in the Polish dataset. To illustrate the process, examples will be discussed here, such as languages, amounts and dates.

***"The best dataset was offered by Denmark, because it complied with all categories and was published in CSV."***

Including data from all European Union member states leads to having to deal with several different languages, since data is often only published in the member state's own language. This is true for both researched funding periods. Therefore, at least the column names had to be translated to get an understanding of the data. We used Google translate for this, when our team did not cover the language themselves. The translation process was quite tedious, because only translating column names does not necessarily yield sufficient information to map the data. Often multiple rows had to be translated in order to assure the column was understood correctly.

Dates can be very difficult from a programming point of view, because they are often formatted very differently e.g. 01/12/2014 and 2014/12/01 and 01. December 2014. Sometimes there were only single years included as dates such as "2014". However, dates can be as detailed as Day/Month/Year. Enabling comparison of dates therefore requires some programming.

# 4.1 AMOUNTS

Amounts are similar to dates, because they require extra programming to get them into the same format. This often concerns the decimal separator and the thousand separator which are usually either 1.000.000,00 or 1,000,00.00. After accounting for these different formats, however, we noticed inconsistencies within the original datasets that made this a very complicated task. The amounts within one dataset had to be unified before all datasets could be brought into the same format. Furthermore, we found examples of numbers that were simply false such as "1 18.245,00€".

Another major issue when wanting to compare amounts between EU countries pertains to the different currencies being used. Amounts listed in the non-euro countries are only listed in the country's own currency, such as with Danish Kronas. While conversion itself is not an issue, it is unclear what date to use for the conversion. The starting date of the project? We simply do not know when the EU transfers the amounts, which highly influences the conversion. Releasing the data in a different currency than Euro is definitely a hindrance in making EU data comparable.

Amounts also differ in their definitions: it is not always clear what a "total amount" is. Does total refer to the entire cost of the project? Or is it simply the sum of both EU financing and national public funding? Where are the third party funds considered in these definitions? Throughout the 2007-2013 data there is no coherence regarding what amounts were published. Some countries publish what they call total amounts, where it is indicated that this amounts consists of the EU's cofinancing amount and the member state's share. Other countries publish only the EU's cofinancing amount, while others (Italy and Sweden) included detailed information on how the member state's share is made up.

After reviewing all the available data, we found that the two most common amounts are a "total amount" indicating the amount financed by the EU + the amount financed by the member state and an "eu cofinancing amount" which indicates the exact amount of EU funding received. However, the first case does not enable calculating how high EU cofinancing is for the respective projects, again making comparisons very difficult.

In order to create a unified dataset, we mapped all the data against our fiscal data model (a list of all variables used is included in the appendix). As discussed, the two most common amount variables are "total amount" and "eu cofinancing amount". Additional variables are "member state amount" if the exact amount a member state paid was indicated and "third party amount" if there was an additional amount indicated paid by any third party (not the member state or the EU). The suffix "eligible" indicates that the amount is not a final amount, but the maximum amount the project is eligible for. This usually applies for the 2014-2020 period, where no final amounts have been declared yet.

# 4.2 EU VARIABLES

Throughout the data several EU specific variables are included that pertain to the disbursement of the funds, however, the terminology used is not coherent. There are multiple terms used that deal with the EU's funding objectives such as category of intervention, theme name, investment priority and priority axis, which seem to be used synonymously. Some of the differences seem to arise between the 2007-2013 and 2014-2020 funding periods.

Therefore, we created four variables to map the information against: theme code and theme name, priority label and priority number. Theme name refers to the EU's objectives (see here), such as "1. Strengthening research, technological development and innovation", while theme code lists the number (1) of the thematic objective. Priority label on the other hand lists the more detailed description of one of the themes such as: "1a Fostering innovation, cooperation, and the development of the knowledge base in rural areas", 1a indicates the priority number. Sometimes both the names and codes are published. More often it is one or the other, requiring all these four variables be present in the data. The term "category of intervention" is more frequently used for the 2014-2020 period and was mapped to priority label. In general, priority label is the more frequent variable in the data, but not nearly frequent enough to allow thorough research.

Additional mapping was required for management authorities, operational programmes and CCI programme codes. Management authority includes information on the administration that supervised the disbursement of the funds. Any column with a similar (translated) meaning as management authority was mapped accordingly. This was done similarly for operational programme, which is a reference to the official document discussing the funding details between the EU and the member state. CCI programme codes were manually included, when we were able to assign them. They can be used to identify the operational programme in case this was not included. CCI codes are assigned per fund (ERDF, ESF, CF), per country/region and sometimes per funding priority, which

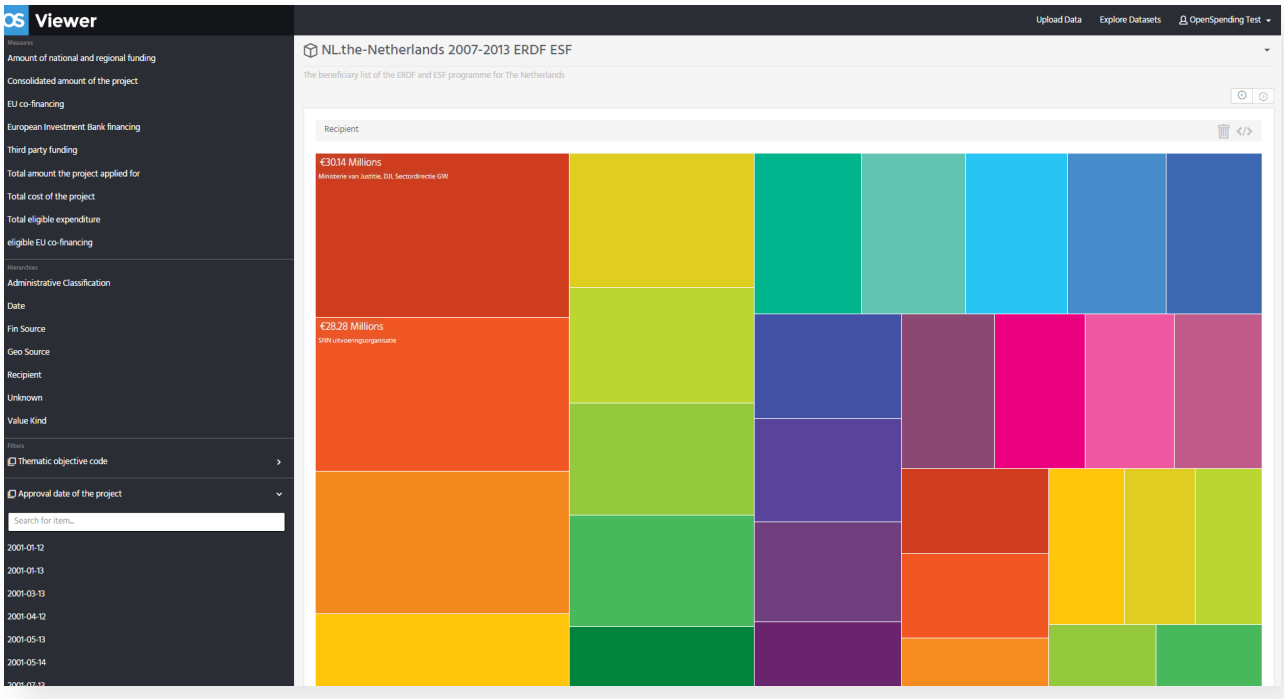


Figure 6.
OpenSpending Treemap Visualisation

makes uniquely identifying them difficult. If we did not have the funding priorities included in the data, assigning CCI codes was not always possible. Furthermore, some projects are funded by multiple funds like (ERDF and ESF) creating unique CCI codes for jointly financed projects. More information on how CCI codes are defined can be found here. Member states should be required to publish their data including a column for CCI codes. This would allow for unique identification of which project was managed by which management authority and governed by which operational programme.

## 4.3 VISUALIZATION

After the data was mapped we uploaded it to Open Spending, where the data can be easily visualized using Open Spending's integrated tools. The picture below shows the Open Spending Viewer where the 2007-2013 Dutch ESIF dataset is loaded. The chosen visualization is a tree map filtered by beneficiaries and showing EU subsidies. The underlying variable is "eu cofinancing amount", showing the exact amount of EU funds distributed to Dutch beneficiaries. Clustering the amounts on beneficiaries enables further selection upon those. One can see which beneficiary received how much money in total and then (upon further selection) see which different projects they executed. This added value is created by OpenSpending which introduced this hierarchy of beneficiary > project to the data and enables viewing the largest beneficiaries per country / region.

A list of the top 30 beneficiaries and their share of the 2007-2013 ESIF funds can be found below. It gives a general idea of what kind of beneficiaries receive co-financing. All of the beneficiaries seem to be related to

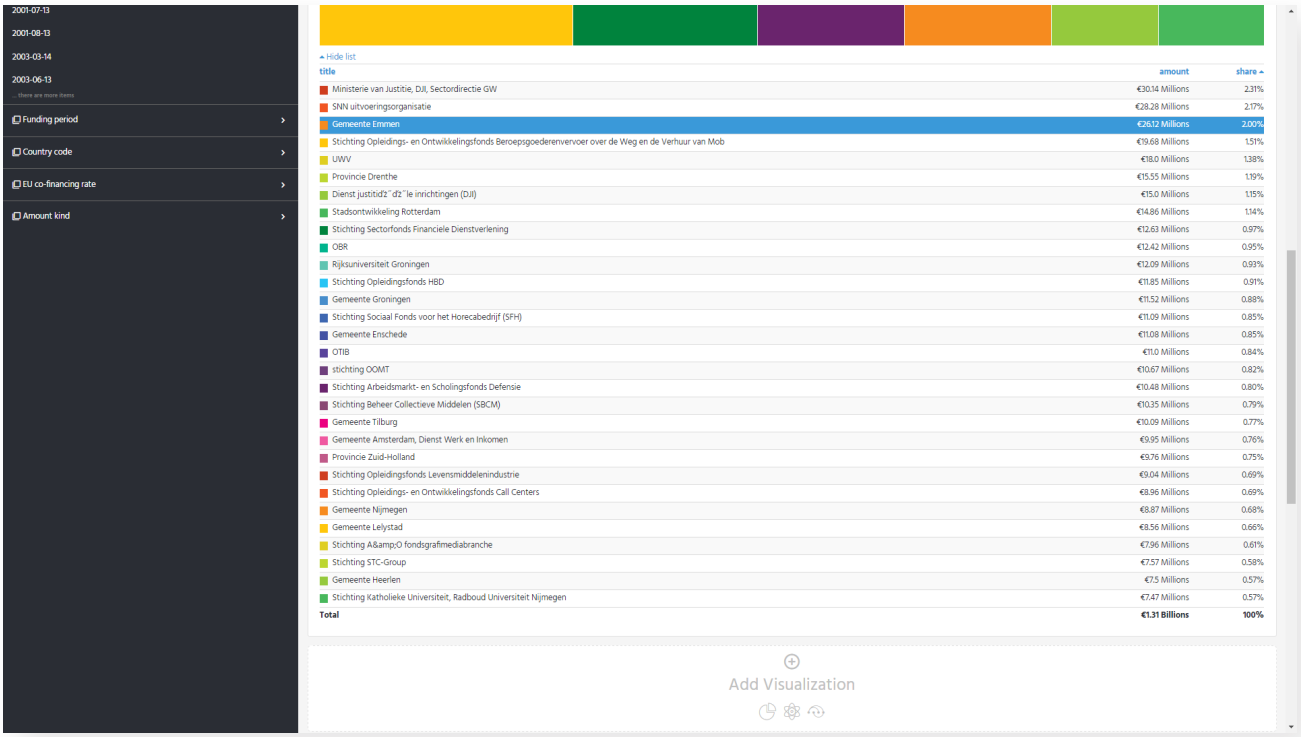


Figure 5.
OpenSpending Beneficiary List

the public sector in some way, showing ministries and counties as large beneficiaries. Furthermore, public universities and foundations are the other main type of institution present in this top 30 list. This pattern of ESIF funds being distributed mostly to public beneficiaries was found to be present throughout all EU member states during our research. However, defining what is a “public” institution is  difficult due to the different legal systems in the EU. Nonetheless, the management authorities are required to publish the beneficiary‘s legal status.

Since the data was mapped to our fiscal data model and we know that we are comparing equal to equal, statistical analysis can be done. Using OpenSpending’s datamine tool, one can use SQL queries to evaluate the data. The graph below analyzes the average EU cofinancing amount spent per project in the 2007-2013 period. This analysis was only possible for those datasets that included the eu cofinancing amount. Additionally, the currency had to be converted to Euros for Poland and the Czech Republic. As discussed before this is not a simple task and we consider this only an approximation. The project’s starting dates were used as the point in time giving us the foreign exchange rate.

# CONCLUSION

This report gave an overview on the quality of the ESIF spending data published by member states in the 2007-2013 and 2014-2020 periods. After summing up the EU's policy background, the member states' data portals were evaluated. We concluded that only 16 of 28 EU member states have an English portal, which makes locating their spending data quite difficult and requires improvement. Furthermore, closed data formats are still common with one PDF and five webapps being used with a total of 23 datasets published so far in the 2014-2020 period. However, comparative analysis showed that substantial progress was made with the introduction of the new Regulation (EU) No 1303/2013 of December 2013. The current funding period shows more machine readable data formats and the data quality has increased. Nonetheless, member states are still slow regarding the data's publication and some not complying with regulatory data publication requirements. Furthermore, issues remain regarding the comparability of amounts, with different currencies and definition of amounts being the most pressing.

Making the received funds comparable should be of the highest priority because it allows for thorough statistical analysis. Including CCI program codes could enable linking the data to the EU's own data portal, uniting spending data with administrative documents such as operational programmes. Furthermore, adding information on the legal form of beneficiaries would improve research opportunities extensively. Lastly, it has to be stressed that only CSV and JSON files can really be considered machine readable and requires adaption.

# APPENDIX I: LIST OF VARIABLES USED IN FISCAL DATA MODEL

| VARIABLE TYPE | DESCRIPTION | NAME |
|---|---|---|
| string | name of the beneficiary (person, company, organisation) | beneficiary_name |
| string | name of project | project_name |
| string | description of the project | project_description |
| numeric | unique code of the project (generated by authority itself) | project_id |
| string | name of person responsible | beneficiary_person |
| string | status of the project | project_status |
| numeric | starting date of the project | starting_date |
| numeric | completion date of the project | completion_date |
| numeric | approval date of the project | approval_date |
| numeric | date on which the final payment was made | final_payment_date |
| string | name of the thematic objective | theme_name |
| numeric | code of the thematic objective | theme_code |
| numeric | CCI codes identifying operational programs | cci_program_code |
| string | description of the priority number of the grant agreement | priority_label |
| numeric | priority number of the grant agreement | priority_number |
| string | management authority | management_authority |
| string | information which operational program the project is governed | operational_programme |
| numeric | total cost of project | total_amount |
| numeric | total eligible expenditure | total_amount_eligible |
| numeric | amount that is awarded from national funds | member_state_amount |
| numeric | amount of co-financing from the EU | eu_cofinancing_amount |
| numeric | amount of co-financing a project is eligible for | eu_cofinancing_amount_eligible |
| numeric | rate (percent) of co-financing from the EU | eu_cofinancing_rate |
| numeric | total amount additional to the action over third party funding | third_party_amount |
| string | acronym of the fund (ERDF, ESF, CF) | fund_acronym |
| string | full address of the beneficiary | beneficiary_address |
| string | city of beneficiary | beneficiary_city |
| string | postal code of beneficiary | beneficiary_postal_code |
| string | region matching the NUTS code | beneficiary_nuts_region |
| numeric | NUTS code of beneficiary region | beneficiary_nuts_code |
| string | county of beneficiary | beneficiary_county |
| string | country of beneficiary | beneficiary_country |
| numeric | two digit NUTS country code of beneficiary | beneficiary_country_code |
| string | URL of the project | beneficiary_url |
| string | a source url of the original data | source |