subsidy stories.eu

# Policy & Data Background

A project by

and

made possible through funding by:

and

http://www.subsidystories.eu

# Contents

In this project the European Regional Development Fund (ERDF), European Social Fund (ESF) and Cohesion fund (CF) data for both the 2007-2013 and 2014-2020 period were collected for all EU member states. 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. This document shall give a brief overview on what the European Structural Investment Funds are and how they work. Furthermore, the process of finding the data and the problems we faced therein will be discussed.

## Governance

The EU Commission laid out their Horizon 2020 strategy for generating smart, sustainable and inclusive growth in the EU. They list detailed investment priorities and thematic objectives (see below). In order to achieve these goals, the EU manages the European Structural Investment Funds, which are the EU's main investment policy tools. The European framework constitutes funding periods of seven years with the last period ranging from 2007-2013, while the current lasts from 2014 until 2020. Institutionally, the member states and the European Commission (through its directorate generals) negotiate a Partnership Agreement within the benchmarks that are set by the regulations for the structural and cohesion funds. Partnership agreements are basically a contract governing the funding process between the European Commission and the member states. Thereafter, operational programmes have to be submitted on how the applicant is planning to achieve the Commission's goals by funding local projects.

The applicant for these operational programmes can be a national or regional institution such as ministries of finance or regional administrations and are referred to as "management authorities". The national / regional distinction arises from a member state's government structure, in countries with strong federal states such as Germany, Spain and Austria applications are based on the member states regions. Thus, funds can be administered on the national or regional level depending on the country. The management authorities have to describe in detail what goals they plan to achieve using the respective ESIF funds and how they will do that. 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 (Austria, Cyprus, Malta, Netherlands, Romania, Spain).

## The funds

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 focus 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 (see). The ESF is Europe's main instrument for supporting jobs and ensuring fairer opportunities for all EU citizens (link). 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: concerns Bulgaria, Croatia, Cyprus, the Czech Republic, Estonia, Greece, Hungary, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovakia and Slovenia.

## Finding EU Data

While the EU provides spending data on the aggregate (member state or regional) level, this project gathered all available data for both the 2007-2013 and 2014-2020 funding periods, on which beneficiaries receive European funding and which projects are implemented. This timeframe was chosen since EU member states have been required to publish the data online starting 2007-2013 period. The management authorities usually created a website regarding the European Structural Investment Funds (ESIF), where they offer information on funding opportunities for potential beneficiaries and list previous projects etc. In some cases, this means there is one website, where information on all funds (ERDF, ESF and CF if applicable) is provided such as France, Cyprus or Denmark. As discussed above in countries with a decentralized state regions are the funds' management authorities and therefore publish the data. In the case of Germany and its 16 regions, this results in a total of 28 websites since a number of regions have decided to create individual websites for individual funds. You can find information on where we found the data on Github (discussed in the manual).

The EU provides an overview on some of the websites in their own portal, available here. It is a good starting point, but not up to date any more. Online searches of "ERDF/ESF beneficiary France" 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 rather challenging.

## Availability

In general, we can say that the data from the 2014-2020 period is substantially better regarding quality and accessibility than the previous period. This is likely due to the fact that 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 a specifically dedicated online portal 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 "not assigned" 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. For now, we are still waiting on some datasets and the version of the data you have is the most complete we can offer. The 2007-2013 funding period is a more difficult obstacle, because the EU regulations were still very vague – only that member states had to publish the funds' beneficiaries and the amounts they received.

## Data Formats

Therefore, the variety of data we encountered varied largely both in format and the information included. Almost half of the data we found was only provided in PDFs, with less than half of the data in more or less open formats such as XLS, XLSX and CSV. Additionally, there were 12 web portals that we had to scrape for the data, for a more detailed overview see the following table.

| | Type of Data (after scraping) | | Type (before scraping) |
|---|---|---|---|
| 0 | WEB | 12 | WEB |
| 89 | CSV | 12 | CSV |
| 0 | XLS | 7 | XLS |
| 0 | XLSX | 28 | XLSX |
| 0 | PDF | 49 | PDF |
| 19 | JSON | 0 | JSON |
| 108 | Total | 108 | Total |

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. In some cases, this is a rather simple task for the newer XLSX format which mostly entails cleaning up headers and overall structure. However, getting the data out of the PDF format is more tedious, since 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 programming if you are an experienced developer or with automated tools such as Tabula and OpenRefine.

Since the share of PDF documents was so large, we chose to use the automated tools when feasible and used our developer only in tricky cases. Tabula is a software that reads a PDF document and tries to detect its structure (in our case mostly tables) – it then gives you some options to select the content from the PDF and extract it into a raw CSV file. This CSV file should include all the data which the PDF included, but is not necessarily structured yet (columns and rows are often unorganized, due to the PDF). In order to restore the data to its original structure, the raw CSV data was then uploaded to OpenRefine, which helps managing large amounts of data. The process of extracting the data from one PDF could take anywhere between 10 minutes to 2 hours, depending on how complicated the PDF was structured.

Another source of data are web portals such as the French 2007-2013 site, which shows a map indicating which region/city/municipality received what amount of funding. While these type of portals are a good way of visualizing data, they do not enable any comparing projects to one another, because single projects have to be selected. Data cannot be aggregated and is difficult to retrieve, because of the way it is embedded in the website. Our developer often spent several hours at a time coding to retrieve the

underlying data.

It is important to note, that since large amounts of the data were not machine-readable and had to be "scraped" we cannot guarantee the completeness of those datasets. It is possible that rows have been dropped in the process although we took every precaution to prevent that. Furthermore, due to the messiness of the certain files, we cannot guarantee that every amount is correctly displayed. Before quoting single amounts for specific projects it is advisable to check the original file.

Some useful links on the EU framework and additional (meta) data for research can be found here:
Data for research
EU Regulation
Categorisation and CCI Codes 2014-2020
Information on themes and cci codes 2007-2013
Wikipedia Structural and Cohesion Funds