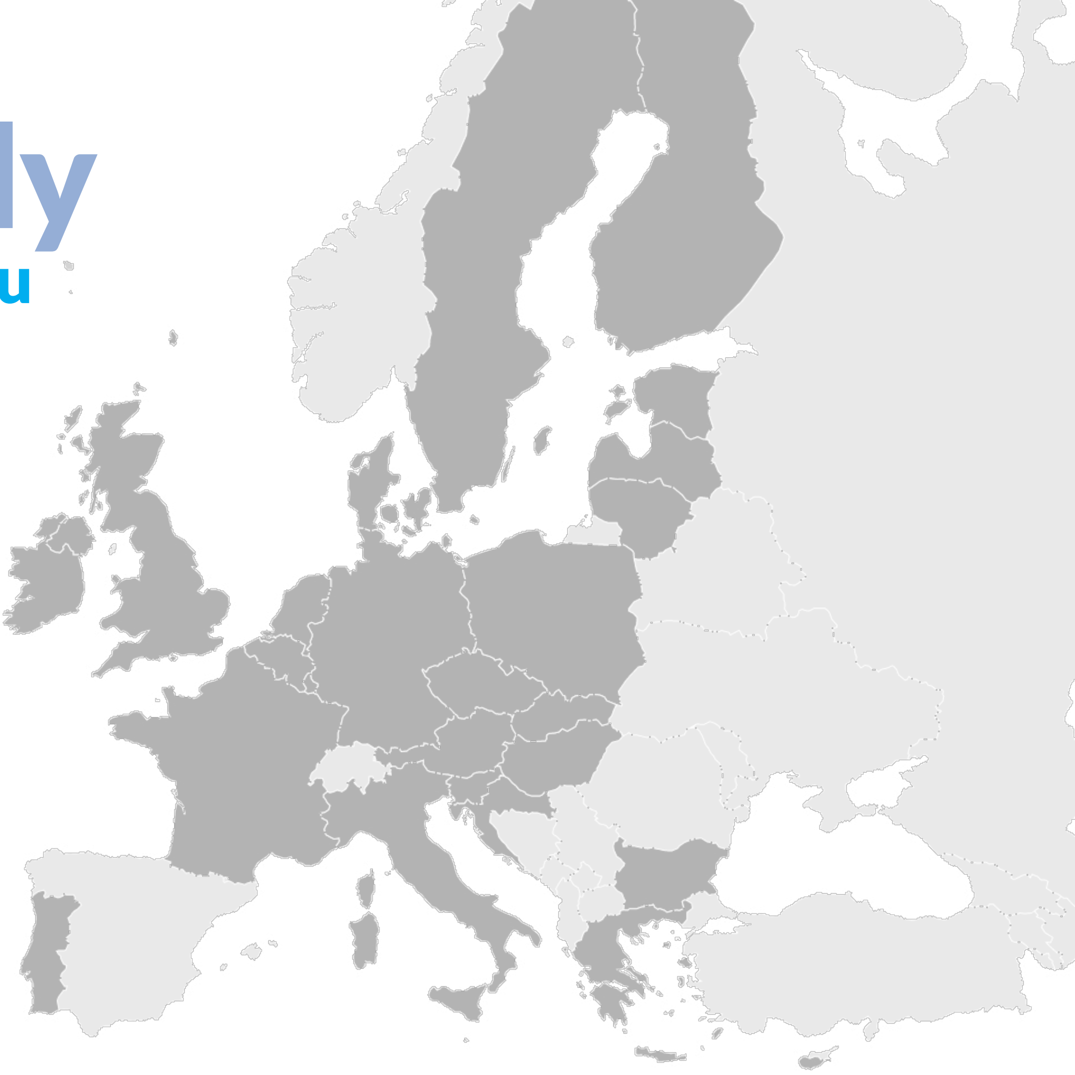
subsidy stories.eu



Methodology & Variables

A project by

OPEN KNOWLEDGE and OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

made possible through funding by:

ADESSIUM FOUNDATION and OpenBudgets.eu

http://www.subsidystories.eu

Contents

This document provides an overview on how we processed the data from its original format into OpenSpending and how to work with it. 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. In order to provide guidance, this documentation explains the decisions we made when trying to unify the national data into one large EU subsidy dataset. Furthermore, it will go into detail on problems we discovered and issues that remain.

The collected data confronted us with two major issues: format (discussed in "obtaining the data") and content. We had to find a common denominator to enable comparing projects across different countries, guaranteeing that, for instance, 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.

Languages

Including data from all European Union member states means dealing with several different languages, since data is often only published in the member state's own language. 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 (German, English, Italian, Spanish, Dutch and French). 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. Nonetheless, translation errors might have been made in the process and we are happy to receive feedback.

Beneficiary Names

Due to varying sources of data input in the original datasets, typos and incorrect spelling may have occurred in the data. This is problematic, because we want to group the same beneficiaries together, assuring that all the projects executed by one beneficiary can be seen. For example "Open Knowledge Foundation" and "Open Knwoledge Foundation" are not spelled the same, and would therefore not be grouped together by OpenSpending / statistical Software. However, clearly the same beneficiary - the Open Knowledge Foundation - was meant to be indicated. Therefore, we used the fingerprints python library to group together similarly spelled beneficiaries. The result of this can be found in the beneficiary_id column, while the words may not be in the right order, beneficiaries are spelled the same way and enable grouping. Furthermore, the beneficiary names displayed on our website and the OpenSpending Viewer are grouped by the beneficiary_id i.e. the fingerprints. This was implemented only for the countries where the fingerprints system worked well: Austria, Estonia, Germany, Italy, Netherlands and United Kingdom. Nonetheless, cases remain where the same beneficiary could not be grouped, because the fingerprints library is not infallible.

Dates

Dates can be problematic from a programming point of view, because they are often formatted very differently (01/12/2009 vs. 2009/12/01 vs. 01. December 2009). This was tackled by using a flexible algorithm allowing for several different date formats, but led to tedious formatting nonetheless, because of incoherencies within datasets. The least information necessary to qualify as a date in our fiscal model is a single year like 2010. However, dates can be as detailed as Day/Month/Year. We distinguish between four different kind of dates common in the data and mapped the data accordingly. *Starting_date* and *completion_date* are in reference to the project actually starting and being finished. *Approval_date* and *final_payment_date* draw on the process of disbursing the funds, approval starts the project, until the final payment is disbursed, ending the project from a fiscal standpoint.

Amounts

Amounts are similar to dates, because they require extra programming to get into the same format. This is often concerned with the decimal separator and the thousand separator which are usually either 1.000.000,00 or 1,000,00.00. Accounting for these different formats is rather simple, however, incoherencies within a dataset made this a very complicated task. Additionally, we implemented an algorithm to spot the numbers formatted differently than the rest of a dataset and corrected these. We try to assure that all amounts are transposed consistently in our database, but since the original data was often flawed, we cannot guarantee that all amounts in our datasets are correct. When a number seems questionable it might be worth comparing it to the original data.

Amounts also differ in their definitions: it is not always clear what a "total amount" is. Is it the entire cost of the project? Or is it simply the sum of EU financing and national public funding? Or could total also include third party amounts? After reviewing all the available data, we found that *total_amount* indicates the amount financed by the EU + the amount financed by the member state. If an amount declared in a dataset is not a total amount, we mapped it to the other options such as: *eu_cofinancing_amount*, the amount paid to the project by the EU or *member_state_amount*, which is the amount the member state contributed to the project. Additionally, a third_party_amount exists, signifying 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. We mapped all the amounts present in the data, so if there is only the *eu_cofinancing_amount* or only the *total_amount* that is due to the original data.

EU co-financing rates

Regarding the rates with which the EU co-finances these projects, we have listed these in the variable *eu_cofinancing_rate* which should be either a decimal number such as 0.5 or 0,5 signifying 50% or an absolute number 50 – also indicating 50%. If this rate is present in the data for every row and the total amount is present as well, you should be able to calculate the eu_cofinancing_amount yourself. However, this is not a frequent case – either detailed information is present and multiple amounts are presented, or co-financing rates are simply not present. The problem with the *eu_cofinancing_rate* is that the EU based them on what kind of funding priority (see below) the project follows. Information on theme names and priority labels are often not included making it impossible to identify the rates. Furthermore, rates may differ within each fund (ERDF, ESF, CF) and each region. The 2014-2020 data is usually better in this case, because it often lists the co-financing rate, enabling the calculation of distinct amounts – usually included already.

Minimal Required Information

Our minimal requirements for uploading datasets are that they include at least beneficiary names, an amount and a date to uphold a working standard. Since not every dataset included a *total_amount* or an *eu_cofinancing_amount* we had to create one unified amount, that is used for illustration. This unified amount variable is called *amount*. Additionally, we created *amount_kind* to indicate which variable included the underlying information. The amount variable follows the hierarchy of total amount > total amount eligible > eu amount > eu amount eligible, because total amount is the most frequent value in all datasets. Therefore, one should never calculate averages using the amount variable, but only the specific amounts, because only they can be easily compared.

EU Specific Variables

Variables with EU specific information is a little difficult, because they are not uniquely identified and multiple long words are more prone to translation errors. Therefore, they will be discussed in more detail here: *management_authority* includes information on the administration that supervised the disbursement of the funds. Any column with a similar name / meaning like 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 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 what CCI codes mean can be found here.

There are multiple variables included dealing with the EU's funding objectives such as *theme_code*, *theme_name, priority_label* and *priority_number*. The terminology within the EU regarding its funding objectives is very difficult, because terms such as category of intervention, theme name and investment priority seem to be used synonymously. Furthermore, there seems to be some difference between the 2007-2013 and 2014-2020 funding periods. *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*. The term "category of intervention" is 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. *Theme_name* was included as well because more detailed datasets listed both. We are happily accepting feedback on this, if mistakes are spotted or better solutions for the mapping are discovered. Further information on the EU's policy background can be found in the specific document.

Regions and Countries

Country names are visible in the meta data, but you can also select the variables *beneficiary_country* and *beneficiary_country_code*. The country code is the two digit NUTS code such as DE for Germany or FR for France. Additionally, regional information is often indicated and was mapped to *beneficiary_nuts_region*, with *beneficiary_nuts_code* included if available. Accordingly, the cities and counties of the beneficiaries were mapped when available, and the same goes for the address (usually street and house number). An overview on NUTS codes can be found on Wikipedia. The country codes provide a good filter variable for SQL analysis.

Datasets and how to use them

To access our datasets follow this link, which leads you to the OpenSpending viewer. You are automatically logged in to our test account and have access to all our data. The joint dataset is called "Complete European ESIF Funds Beneficiaries 2000-2020" and includes all data. Furthermore, all single country/ region datasets are accessible there too. To see an overview on what datasets are available go to the top right corner and click on OpenSpending Test (this is our test account) and then select "profile". This page offers an overview on the latest version of datasets available.

An overview on all mappings can be found below, showing variable names and a brief description. Variables can be clustered into two main types: string or numeric. A numeric variable is based on numbers such as dates or amounts. A *string* means that the content is text based (words).

Name	Description	Variable Type
beneficiary_name	name of the beneficiary (person, company, organisation)	string
beneficiary_id	fingerprint of the beneficiary_name for easier comparison	string
project_name	name of project	string
project_description	description of the project	string
project_id	unique code of the project (generated by authority itself)	numeric
beneficiary_person	name of person responsible	string
project_status	status of the project	string
starting_date	starting date of the project	numeric
completion_date	completion date of the project	numeric
approval_date	approval date of the project	numeric
final_payment_date	date on which the final payment was made	numeric
theme_name	name of the thematic objective	string
theme_code	code of the thematic objective	numeric
cci_program_code	CCI codes identifying operational programs	numeric
priority_label	description of the priority number of the grant agreement	string
priority_number	priority number of the grant agreement	numeric
management_authority	management authority	string
operational_programme	information which operational program the project is governed	string
total_amount	total cost of project	numeric
total_amount_eligible	total eligible expenditure	numeric
member_state_amount	amount that is awarded from national funds	numeric
eu_cofinancing_amount	amount of co-financing from the EU	numeric
eu_cofinancing_amount_eligible	amount of co-financing a project is eligible for	numeric
eu_cofinancing_rate	rate (percent) of co-financing from the EU	numeric
third_party_amount	total amount additional to the action over third party funding	numeric
fund_acronym	acronym of the fund (ERDF, ESF, CF)	string
beneficiary_address	full address of the beneficiary	string
beneficiary_city	city of beneficiary	string
beneficiary_postal_code	postal code of beneficiary	string

beneficiary_nuts_region	region matching the NUTS code	string
beneficiary_nuts_code	NUTS code of beneficiary region	numeric
beneficiary_county	county of beneficiary	string
beneficiary_country	country of beneficiary	string
beneficiary_country_code	two digit NUTS country code of beneficiary	numeric
beneficiary_url	URL of the project	string
source	a source url of the original data	string