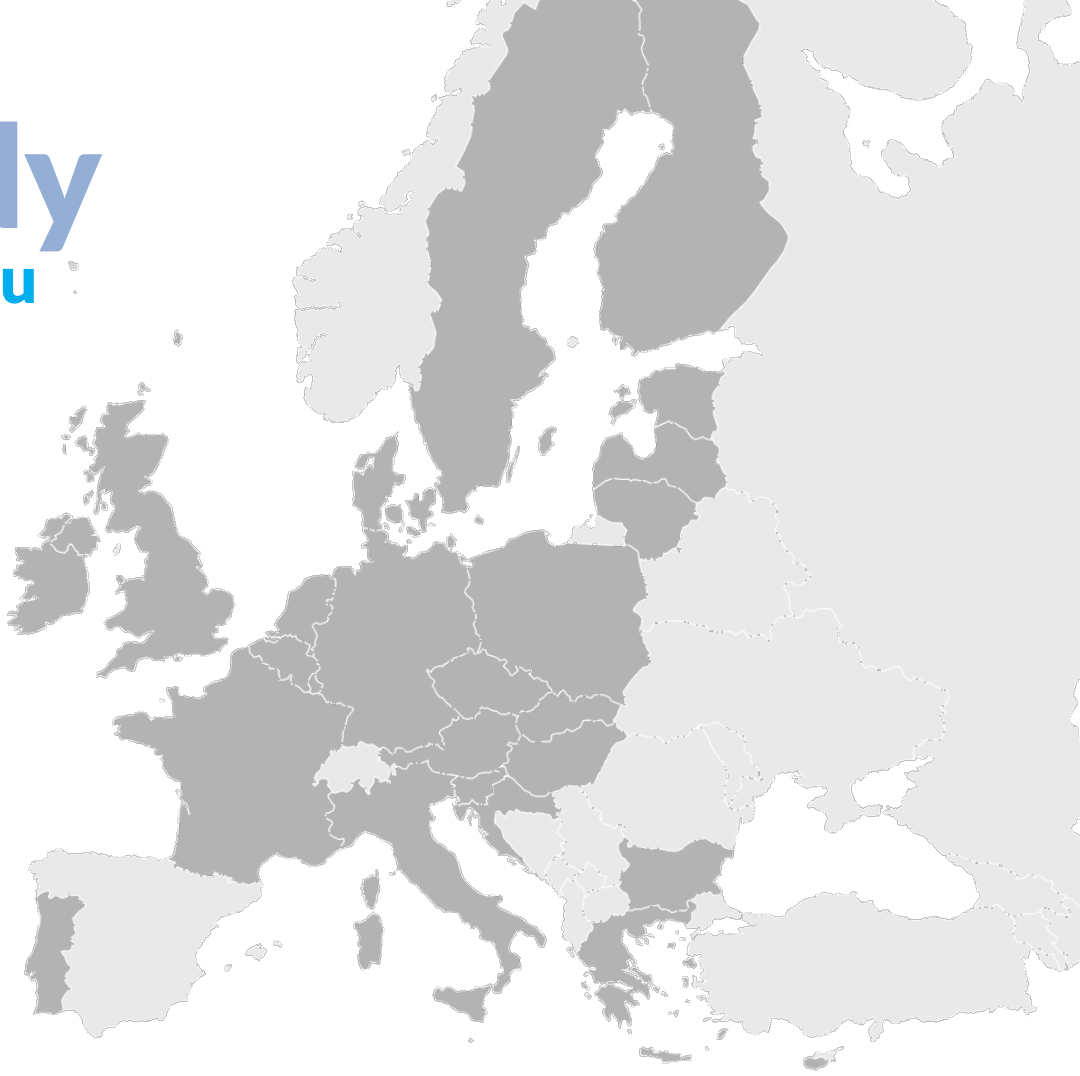
subsidy stories.eu



Dataset Descriptions

A project by

and

made possible through funding by:

ADESSIUM FOUNDATION

http://www.subsidystories.eu

Contents

Note

This is an overview of the original files and transformations that have gone into the full dataset, that is used for subsidystories.eu. If you need detailed information about the original datasets, descriptions, and transformation executed and the original portals, this is all available on github. In our github, go to Data, country name, optional region, fund name and period, where you find the description file - detailing the finding place, column names etc. - and the original datafiles themselves.

Concatenated Dataset

The concatenated dataset is a large version, which combines the datasets of all 28 EU member state countries. The data is collected from national and regional portals of the member states where the beneficiary lists are published. We have collected 115 datasets, 21 datasets are missing. This means that for most countries, we have identified all data for all Structural Funds. However for Ireland, UK, Spain, and Austria we still have major gaps.

For creating the joined datasets out of these files, we mapping every single datasets to common denominator variables that can be compared between the countries. You can find a list of all variables used below. More information on the methods can be found in the Methodology section.

Name	Description	Type
beneficiary_name	name of the beneficiary (person, company, organisation)	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 3rd 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

Austria

The dataset for Austria contains ESF and ERDF Beneficiary lists for the period 2007 - 2013.

ESF 2007-2013	Österreich	
ERDF 2014-2020	Österreich	missing
ERDF 2007-2013	Burgenland	
ERDF 2007-2013	Niederösterreich	
ERDF 2007-2013	Kärnten	
ERDF 2007-2013	Steiermark	
ERDF 2007-2013	Oberösterreich	
ERDF 2007-2013	Salzburg	
ERDF 2007-2013	Tirol	
ERDF 2007-2013	Vorarlberg / Austria	

The data for the funding period 2014 - 2020 for both ERDF and ESF has been requested, but was not available by January 2017.

The dataset for Austria is built up from 9 single datasets, published in PDF. The European Social Fund Dataset is published on a national portal, the datasets for ERDF Beneficiaries were published on 8 regional portals. For the original files + portal website - see data on github.

The datasets contains the basics columns: beneficiary name, operation name, total amounts, whether it is only requested or already paid. We could additionally assign CCI code, region, funding period, and fund name.

Belgium

The dataset for Belgium is - like the country - built up of very different regional datasets:

Region Brussels	ERDF 2007-2013
Région Wallonne / Brussel Capitale	ESF and ERDF 2014-2020
Flanders	ERDF 2007-2013
Flanders	ERDF 2014-2020
Flanders	ESF 2007 - 2013 and 2014 - 2020.

As you can see here, the fact that Wallonia and Brussels were not presented together in the period 2007 - 2013, but were in 2014 - 2020 made it impossible for us to determine whether we indeed had the Wallonian Regions dataset. After several searches, we now set the 2007 - 2013 data for Wallonia as missing. It also must be noted that the 2014-2020 datasets still had very little projects.

The dataset for Belgium is built up from the five above datasets (originals can be found on github). Brussels/Wallonia is scraped from its website, and Flanders could be downloaded in csv/excel. Flanders 2014-2020 contains rich beneficiary information and co financing information, also current brussels/wallonia datasets does contain priority numbers, additional information on co-financing etc.

Bulgaria

The Bulgarian data could be directly downloaded from their portal. However, there were some issues with the file format, which we managed to resolve. The data is very basic in nature but covers both

the 2007-2013 and the 2014-2020 period. It does not include any information on dates so distinction between periods is not possible. Amounts are presented in the Bulgarian currency LEV.

Croatia

For Croatia all data was available from the portal for the entire funding period (from 2008 onwards) for all three funds: CF, ERDF, ESF.

The data still had to be converted from XLS to CSV to provide the right descriptions, date formats have been adjusted, currency conversions to Euro have been done for the additional amount column, and the funding period is based on the application date column.

The data is rich, the priorities and program are indicated per project, additional information is provided, and co-financing rates, application and completion dates are available etc.

Cyprus

For Cyprus we only have one dataset that contains ERDF and Cohesion Fund Beneficiaries 2007 - 2013. The ESF fund for both periods was indicated on the website, but not available in January 2017.

The raw data could be easily downloaded from the portal, but needed extensive cleaning as the excel file itself contained extensive non machine readable layouts, finally we needed to clean up oddly formatted dates.

In terms of available information, the dataset is basic. It contains beneficiary name, project names, dates and amounts, fund and priority axis. We have added funding period and region.

Czech Republic

For the Czech Republic, the ERDF data could be downloaded in Excel format, making it fairly easy to work with. Datasets are present for both the 2007-2013 and the 2014-2020 period, however, the ESF data could not be found. The amounts are in Czech Krona, which has to be considered (only the "amount" column is converted as extra info). Data is very detailed offering several different amounts and regional information.

Denmark

For Denmark, we have funding data for both funding periods 2007 - 2013 and 2014 - 2020. The datasets contain both funds: ERDF and ESF.

The data was downloaded from the portal in csv files, but needed conversion and cleaning for dates and amounts. We have added a converted amount column with the amount variable to display the converted amounts, without losing the original data.

The data is extensive in project information, beneficiary information, beneficiary address information, priority axis, co-financing rates, and different amounts applied for. In the transformation of the data, nuts codes have not been drilled down to region. This can be done on the dataset level itself because of

the extensive address data available.

Estonia

For Estonia, we have the data for both funding periods and all funds in one dataset.

The data was scraped from the web portal with a python scraper. We are missing CCI codes and the fund names, which makes it impossible to distinguish ERDF, ESF or CF projects. We have added funding period and country NUTS-code, drill down to regional level is missing.

The data itself is basic, it provides information on the project name and beneficiaries, and additionally co-financing amounts, paid out amounts, management authority, and priority axes.

Finland

For Finland, data for ERDF and ESF for both funding periods is available. The data could be downloaded in CSV directly from the portal. However, it always resulted in a faulty CSV and our developer had to spend a long time fixing the file.

The data contains basic information on projects and beneficiaries, management authority, and is slightly more extensive in terms of amounts, breaking total amounts to EU amount, and member state amount.

France

For France, we found two datasets: 2007 - 2013 and 2014 - 2020. Both datasets contain the beneficiaries and projects for ESF and ERDF in France for all regions.

The original files came in CSV, and XLS, the latter needed pre-processing. Both needed small cleaning actions for dates and amounts and we have added funding periods.

The data is rich, it provides beneficiary name, project and priority, and additional geographic information on the place of the project. Amounts are split up in eu- and national-financing amounts and total amounts. Dates are provided for the start and the completion of the project.

Germany

German Datasets were our biggest headache, they are published on state-level, separated per fund and period, in PDF displaying only the bare minimum of information. In the table you can find the overview of all datasets, missing data etc.

NUTS	State	Fund and Period	Convert	Raw
DE1	Baden-Württemberg	ERDF 2007-2013	CSV	PDF
DE1	Baden-Württemberg	ESF 2007-2013	CSV	PDF
DE1	Baden-Württemberg	ERDF 2014-2020	CSV	XLSX
DE1	Baden-Württemberg	ESF 2014-2020	CSV	XLSX
DE2	Bayern	ERDF 2007-2013	CSV	PDF
DE2	Bayern	ESF 2007-2013	CSV	PDF
DE2	Bayern	ERDF 2014-2020	CSV	XLSX
DE2	Bayern	ESF 2014-2020	CSV	XLSX

DE3	Berlin	ERDF 2007-2013	CSV	PDF
DE3	Berlin	ESF 2007-2013	CSV	PDF
DE3	Berlin	ERDF 2014-2020	CSV	XLSX
DE3	Berlin	ESF 2014-2020	CSV	XLSX
DE4	Brandenburg	ERDF 2007-2013	CSV	PDF
DE4	Brandenburg	ESF 2007-2013	CSV	PDF
DE4	Brandenburg	ESF 2014-2020	JSON	XLSX
DE4	Brandenburg	ERDF 2014-2020	CSV	XLSX
DE5	Bremen	ERDF 2007-2013	CSV	PDF
DE5	Bremen	ESF 2007-2013	CSV	PDF
DE5	Bremen	ERDF 2014-2020	XLSX	XLSX
DE5	Bremen	ESF 2014-2020		missing
DE6	Hamburg	ERDF 2007-2013	CSV	PDF
DE6	Hamburg	ESF 2007-2013	CSV	PDF
DE6	Hamburg	ESF 2014-2020	CSV	XLSX
DE6	Hamburg	ERDF 2014-2020		missing
DE7	Hessen	ERDF 2007-2013	CSV	PDF
DE7	Hessen	ESF 2007-2013	CSV	PDF
DE7	Hessen	ESF 2014-2020	CSV	XLSX
DE7	Hessen	ERDF 2014-2020		missing
DE8	Mecklenburg-Vorpommern	ERDF 2007-2013	CSV	PDF
DE8	Mecklenburg-Vorpommern	ESF 2007-2013	CSV	PDF
DE8	Mecklenburg-Vorpommern	ERDF 2014-2020	CSV	XLSX
DE8	Mecklenburg-Vorpommern	ESF 2014-2020	CSV	XLSX
DE9	Niedersachsen	ERDF 2007-2013	CSV	PDF
DE9	Niedersachsen	ERDF 2007-2013	CSV	PDF
DE9	Niedersachsen	ESF 2007-2013	CSV	PDF
DE9	Niedersachsen	ESF 2007-2013	CSV	PDF
DE9	Niedersachsen	ERDF 2014-2020	CSV	XLSX
DE9	Niedersachsen	ESF 2014-2020	CSV	XLSX
DEA	NRW	ESF 2007-2013	CSV	PDF
DEA	NRW	ERDF 2007-2013	JSON	WEB
DEA	NRW	ESF 2014-2020	CSV	XLSX
DEA	NRW	ERDF 2014-2020	CSV	XLSX
DEB	Rheinland-Pfalz	ERDF 2007-2013	CSV	PDF
DEB	Rheinland-Pfalz	ESF 2007-2013	JSON	PDF
DEB	Rheinland-Pfalz	ERDF 2014-2020	CSV	XLSX
DEB	Rheinland-Pfalz	ESF 2014-2020	CSV	XLSX
DEC	Saarland	ERDF 2007-2013	CSV	PDF
DEC	Saarland	ESF 2007-2013	JSON	PDF
DEC	Saarland	ESF 2014-2020	CSV	XLSX
DEC	Saarland	ERDF 2014-2020		missing
DED	Sachsen	ESF 2014-2020	CSV	XLSX
DED	Sachsen	ERDF 2014-2020	CSV	XLSX
DED	Sachsen	ESF 2007-2013	CSV	PDF
DED	Sachsen	ERDF 2007-2013	JSON	PDF
DEE	Sachsen-Anhalt	ERDF 2007-2013	CSV	PDF
DEE	Sachsen-Anhalt	ESF 2007-2013	CSV	PDF
DEE	Sachsen-Anhalt	ERDF 2014-2020		missing
DEE	Sachsen-Anhalt	ESF 2014-2020		missing
DEF	Schleswig-Holstein	ERDF 2007-2013	JSON	PDF
DEF	Schleswig-Holstein	ESF 2007-2013	JSON	PDF
DEF	Schleswig-Holstein	ESF 2014-2020	CSV	XSLX
DEF	Schleswig-Holstein	ERDF 2014-2020	CSV	XSLX
DEG	Thüringen	ERDF 2014-2020	CSV	XLSX
DEG	Thüringen	ERDF 2007-2013	CSV	PDF
DEG	Thüringen	ESF 2007-2013	JSON	PDF
DEG	Thüringen	ESF 2014-2020	CSV	XLSX

As is visible in the table, almost all data files had to be scraped or converted from PDF to csv/json, or from XLSX to csv. First scraping is always imperfect, so in many files we had to do small cleaning ac-

tions. Major cleaning needed to be done in the field of duplicates. The scraping software and scripts used had the problem that they had produced a number of duplicates. We have then removed these duplicates, however we cannot guarantee that all duplicates have been removed .
Second, dates needed to be adjusted to formats that were recognised by our programs.
Third, the decimal delimiters caused problems, the software usually automatically recognises the delimiter, but in the wide variety of german cases there were odd discrepancies on row level that produced odd outcomes. Through several scripts, we could solve most of these problems afterwards. Finally, we have added columns for fund names, CCI-codes were included, funding period, fund, region.

The german datasets for 2007 - 2013 are basic. Only the following columns are provided: beneficiary, project names, a year, the amount applied for, or the awarded awarded. In the cycle 2014 - 2020 it has improved adding project description, priority axes, and amounts for co-financing etc.

Greece

The greek datasets contain the funds ERDF, CF, and ERDF.
The 2007-2013 data had to be scraped from a web portal, while the 2014-2020 data could be downloaded in CSV.
We have added nuts-codes and funding periods.
The data is rich, it contains the basic information and additional information on IDs, detailed amounts for the EU and the member state amount, priority numbers and themes.

Hungary

The data for Hungary had to be scraped from the websites and is available even for 2000 - 2006. We have modelled and imported all three datasets - by exception: 2000 - 2006, 2007 - 2013, 2014 - 2020. Hungary receives funds from the Cohesion Fund, ERDF, and ESF.

They can be downloaded from the portals of the hungarian government, but downloading all the data as we planned took forever which is why we decide to scrape it.

The data is rich, it contains all the basic information and additional beneficiary information on place, program information, additional amounts, and 2 dates - start and approval. The amounts included are in Hungarian Forint, only the amount column was converted to Euro for orientation.

Ireland

Ireland was one of the hardest datasets to gather, and we are still lacking all data for 2014 - 2020 and all ESF data for both periods. After repeated inquiry, we learned that the data for 2014 - 2020 is not produced yet.

IE	Ireland - southern and eastern	ERDF 2007-2013	JSON	WEB
IE	Ireland - Border Midland and Western	ERDF 2014-2020	Missing	
IE	Ireland - Southern and Eastern	ERDF 2014-2020	Missing	
IE	Ireland National	ESF 2014-2020	Missing	
IE	Ireland National	ESF 2007-2013	Missing	
IE	Ireland North West	ERDF 2007-2013	JSON	WEB
IE	Ireland - Border Midland and Western	ERDF 2007-2013	Missing	

For the three files that are available, we have scraped them from the web portals. As usual, we have added funding period, region nuts code, and fund acronym.

The Southern and Eastern dataset has detailed information concerning project location, amounts, dates, programs and priorities. The Ireland North West dataset in turn is of low quality, it only has total amounts, starting date, beneficiary name and project name.

Latvia

We have excluded the Latvian datasets for 2007-2013, because they did not contain any amounts, which we consider the bare minimum for uploading them.

The dataset for Latvia for the funding period 2014-2020 contains CF, ERDF, and ESF beneficiaries. We have added the funding period column, and the nuts code on country level. The exact fund name is unknown. The dataset itself is limited, it only contains project name, beneficiary, and basic information on amount, and dates.

Lithuania

Lithuania's datasets present ERDF, CF, and ESF all in one file for each funding period: 2007-2013 and 2014-2020. The datasets are presented in an online portal where they could be downloaded in csv files. We have as usual added the Nuts codes and the funding region. This dataset does not have detailed geographic information and was published on national level, so the information concerning regional allocation is missing.
The datasets themselves provide an average detail level. They provide information about the contract status, but no additional project- or beneficiary information.

Luxembourg

The data for beneficiaries of ESF and ERDF in Luxembourg can be found in their web portal, but had to be scraped.
We added NUTS acronym for the national level, but the regional level nuts code per project was not available. Also the funding period is unknown, as the portal provides the datasets for both periods together.
The datasets itself are rich. They contain beneficiary name and address, the project name, program, priority, and theme, start and completion dates, and total, eligible, and co-financing amount and even the EU co-financing amount.

Malta

The 2007-2013 data for Malta had to be scraped from their website. However, it is very detailed and offers distinctions for all funds and various variables such as management authorities, investment priorities and project descriptions.

The 2014-2020 data was only recently released, but only in PDF format making the scraping a difficult task, which is why it is not included in the current iteration.

Netherlands

The Netherlands dataset contains both periods (2007 - 2013) and all structural funds: ESF, ERDF, but also EARDF, and MFF. We have added the funding period, and the national nuts-code.

The dataset itself is very rich. It contains project name and description, beneficiary names, addresses, website, and even co-beneficiaries. It has four amount columns: national, eu, third party, and total amount, and co-financing rate. Finally, the date columns contain dates for start, finalisation, and last updated.

Note: many fields for columns such as co-beneficiaries applicant, website, etc. were empty.
The 2014-2020 was not published as of January 2017.

Poland

The data for the ESF ERDF and CF for both funding periods is available on a web portal, where one can easily download all the data in CSV. We have downloaded it in three CSVs (per fund) for each funding periods.

We have added the Nuts Code for the national level, funding period and funding acronym. We excluded some columns from the mapping that were blank or we could not link to the overall dataset: multimedia, field, and the question whether the project was implemented in more regions.

The remaining dataset is limited, only beneficiary name, project, total and eu-cofinancing amount are given as columns in the original data. The amounts are in Polish Zloty and only the amount column was converted to Euro for orientation.

Portugal

The datasets for portugal are divided by funding period, but contain all the funds: ERDF, ESF, and CF (2007 - 2013 and 2014 - 2020).

They can be downloaded from the portal as xls files, but we have turned them into json files for ease of processing. We have only added PT for the country nuts code.

The dataset itself is very rich. Beyond the basic fields (project, beneficiary, date, and total amount) It contains funds, theme, program, priority, additional address data and additional amounts.

Romania

For Romania, we could not find the right files and funds. We are still in contact with the national administration and waiting for a response.

Slovakia

The data for slovakia had to be scraped from a platform and then extensively cleaned. It contains data

for 2007 - 2013 and 2014 - 2020, ERDF, ESF, CF.

The data was scraped and thus needed extensive cleaning, we also could not map all the columns as not all of them had a clear meaning. We have added columns for NUTS Code for the country and funding period.

The data itself contains the basic columns - beneficiary, project, total amount, date, fund - and additional amounts for national and eu level, start and final date, and project status.

Slovenia

The data for Slovenia comes from a web portal and had to be scraped. We turned it into two json files for 2007-2013 and 2014 - 2020 containing all funds: ESF, ERDF, CF.

We added funding period and nuts code for the country level to the dataset. Due to the scraping, we also had to do rigorous cleaning. All columns could be mapped except for project status, statistical region, and instrument.

The dataset is rich. Beneficiary name, address, url etc. are given. Projects are also described with name and summary, and linked to priority axes, program etc. Finally different dates are given, and the amounts are detailed in member state amount, eu amount, total amount etc.

Spain

The Spanish data is only available for the 2007-2013 period so far. We have only included the ERDF datasets, those were made up of regional PDFs that had to be scraped. Information is rather scarce offering only beneficiaries, projects, dates and amounts. The ESF dataset is on the platform but currently not complete - it also consists of regional PDFs and will hopefully be added soon.

The 2014-2020 data has not been released as of February 2017.

ES	Spain	ERDF 2014-2020	Missing	
ES	Spain	ESF 2014-2020	Incomplete	
ES	Spain	ESF 2007-2013	JSON	PDF
ES	Spain	ERDF + Cohesion 2007-2013	JSON	PDF

Sweden

Both Swedish datasets for 2007-2013 and 2014-2020 had to be scraped from a specific webapp. While the data is detailed on a certain level offering project descriptions, responsible persons and investment priorities, it does not allow for distinguishing between the ERDF and the ESF. The amounts are displayed in Swedish Krona and only the amount column was converted to Euros.

United Kingdom

The UK data for 2007-2013 was quite a mess, because it consists only of PDFs divided per country (Wales, Northern Ireland, Scotland, England) and region. Luckily @Balkey (on Github) had scraped all the data and created a CSV that includes very rich information (eu co-financing amounts, geocodes,

and coordinates).

The 2014-2020 data was collected by us and so far covers all funds for England, Northern Ireland and Wales. Scottish data is only available in PDF form and does not offer sufficient information and was therefore not included. Amounts for all datasets are in British pounds (GBP) - again only the amount column was converted to give some orientation. Amounts for both periods are in British pounds (GBP).

80	UK	England	ERDF ESF 2007-2013	CSV	PDF
80	UK	Northern Ireland	ERDF ESF 2007-2013	CSV	PDF
80	UK	Scotland	ERDF ESF 2007-2013	CSV	PDF
80	UK	Wales	ERDF ESF 2007-2013	CSV	PDF
80	UK	United Kingdom total	ERDF ESF 2007-2013	CSV	PDF
80	UK	City of London	ERDF ESF 2014-2020	CSV	XLSX
80	UK	England	ERDF ESF 2014-2020	CSV	CSV
80	UK	Northern Ireland	ERDF ESF 2014-2020	CSV	XLSX
80	UK	Scotland	ERDF ESF 2014-2020	request	PDF
80	UK	Wales	ERDF ESF 2014-2020	CSV	XLSX