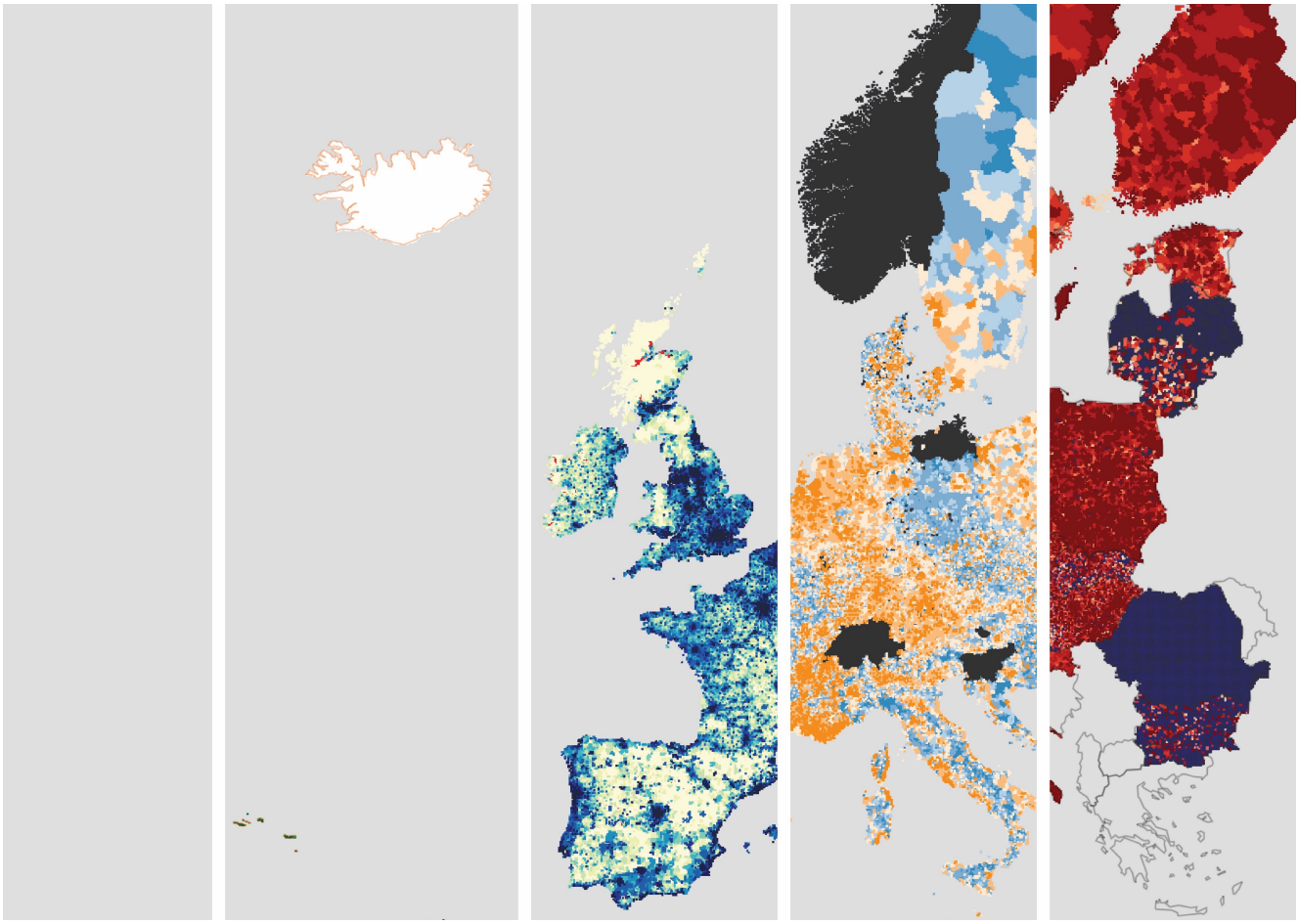
# Progress Report

## The locally funded Europe: analyzing the composition and grant-effectiveness of EU subsidies on a municipality level

Balázs Krich, Aaron Swartz Fellow, 2017
Balázs Bónis
June 27, 2017



# Introduction

With the aim to “promote an overall harmonious development and strengthen economic and social cohesion by reducing development disparities between the regions”[1] the European Union distributed 319.59 billion euros among its member states through the European Strategic and Investment Funds (ESIF) between 2007 and 2013.

Since the Commission Regulation requires that “the list of beneficiaries, the names of the operations and the amount of public funding allocated to operations should be published, electronically or otherwise”[2], in theory it is possible to analyze the utilization of EU funds on a transaction level, encompassing all member states.

However, national governments and local authorities allocating the funding and administering the process did not have further agreements on how to collect and store data, therefore what is published is unstructured, not machine readable and sparse.

So far only SubsidyStories - a joint collaboration between the teams of Open Knowledge Germany and Open Knowledge International (OKI) - tried to compile “a consolidated overview of all the available sources and the distribution of the money down to the transactional level”. According to the owners of the project, their initiative “is unique for it unifies the available datasets of fund distribution on one website”.

As the Aaron Swartz fellow at Open Society Archives (OSA) for 2017, my research proposal was based on the assumption that I will be able to work with the data provided by SubsidyStories, even though it did not hold any kind of geolocational information about the transactions (either about the location of the beneficiary or the location where the supported project was realized).

After teaming up with Balázs Bónis to accelerate the research process, our aim was to come up with different methods to obtain geolocation for each transaction, resulting in two unprecedented, novel analytical observations:

1. Mapping the distribution of funds across the 114,177 level 2 Local Administrative Units (LAU2) - the lowest level components of the Classification of Territorial Units for Statistics (NUTS) regions maintained by Eurostat[3].

2. Pairing external datasets sourced from Eurostat - and possibly other sources - to the subsidy dataset on LAU2 and NUTS3 level, making it possible to determine such indicators as the SUM of subsidies per capita received in each municipality across the EU 28 member states.

---

[1] Council Regulation (EC) No 1083/2006 of 11 July 2006

[2] Council Regulation (EC) No 1828/2006 of 8 December 2006

[3] Greece is not included currently in our dataset, as data could not be attained so far. I still refer to EU 28 in our report as I intend to include Greece in the future.

# Collecting data from external sources

Before working with the transactions themselves, we collected a number of external datasets which were necessary for further analysis - either as tools in creating secondary metrics or as complementary data.

### 1. Correspondence tables between the various levels of the NUTS system

The NUTS system contains the following aggregation levels:

| AGGREGATION LEVEL | NUMBER OF UNITS WITHIN THE EU 28 (2011) |
|---|---|
| NUTS1 (~STATES) | 109 |
| NUTS2 (~REGIONS) | 314 |
| NUTS3 (~PROVINCES) | 1,433 |
| LAU1 (~COUNTIES / DISTRICTS) | 8,772 |
| LAU2 (~MUNICIPALITIES) | 114,177 |

*Table 1: number of units in the NUTS system*

Though Eurostat collects and publishes correspondence tables between the NUTS3 and LAU2 levels, and updates changes from 2010 onwards, translating between other statistical scales was not possible based on these tables. Therefore the NUTS1, NUTS2 and LAU1 codes for each unit had to be collected - typically from national statistical offices across all member states - and more importantly, the member relationships between scale levels - which unit is part of which parent unit - had to be recreated. Mapping these relationships proved to be an essential tool later on in aggregating data attained at the lowest, LAU2 level to higher level scales. Since these boundaries change over time, we decided to lock the geometry at the state recorded at the time of the 2011 EU census, since this was the state which could be connected to most other datasets.

### 2. Geometries for the administrative boundaries of all LAU2 and NUTS3 level units

Geometry files are the basis of all Geographical Information Systems (GIS) - by presenting a set of boundaries in a two dimensional space, we are able to determine if a geolocation - essentially a pair of coordinates - falls inside or outside of the boundaries of a given geometry object, thus we can decide if the data examined at the referred coordinates is a member of a set of aggregates at different levels.

Eurostat offers these geometries, though they are not harmonized with the codes of the NUTS system - meaning a 100 % pairing was not obtainable without corrections. Regarding the set of administrative units at each NUTS level, we decided to accept only those, which we

could identify with a geometry - e.g. if a LAU2 unit was identified at a population dataset, but could not be paired with a geometry, it was discarded.

| COUNTRY | NUMBER OF LAU2 GEOMETRY COLLECTED |
|---|---|
| AUSTRIA | 2,357 |
| BELGIUM | 589 |
| BULGARIA | 5,302 |
| CROATIA | 556 |
| CYPRUS | 615 |
| CZECH REPUBLIC | 6,251 |
| DENMARK | 2,172 |
| ESTONIA | 226 |
| FINLAND | 336 |
| FRANCE | 36,678 |
| GERMANY | 11,413 |
| GREECE | N/A |
| HUNGARY | 3,154 |
| IRELAND | 3,409 |
| ITALY | 8,092 |
| LATVIA | 119 |
| LITHUANIA | 540 |
| LUXEMBOURG | 116 |
| MALTA | 68 |
| NETHERLANDS | 418 |
| POLAND | 2,479 |
| PORTUGAL | 4,260 |
| ROMANIA | 3,181 |
| SLOVAKIA | 2,927 |
| SLOVENIA | 210 |
| SPAIN | 8,116 |

| COUNTRY | NUMBER OF LAU2 GEOMETRY COLLECTED |
|---|---|
| SWEDEN | 290 |
| UNITED KINGDOM | 10,303 |
| **EU 27** | **114,177** |

*Table 2: number of identified LAU2 units per country*

### 3. Territory and population data over time

Also offered by Eurostat, the territory of each boundary unit in $km^2$ and their population was paired to all LAU2 level units, and aggregated to each level of the NUTS system. Population data was collected for the following historical dates:

- 1961
- 1971
- 1981
- 1991
- 2001
- 2011

As a result, we were able to calculate population density and population change over time for the period between 1961 - 2011 for each unit at each level. Since Eurostat also offers yearly population datasets from 2011 onwards, we plan to include these datasets as well in our research. The outcome of the use of this dataset could be to examine if the European Union is funding municipalities with shrinking or growing populations. The short term impact of the ESIF 2007-2013 could also be addressed: does the amount of funding show any correlations in tendencies of population change?

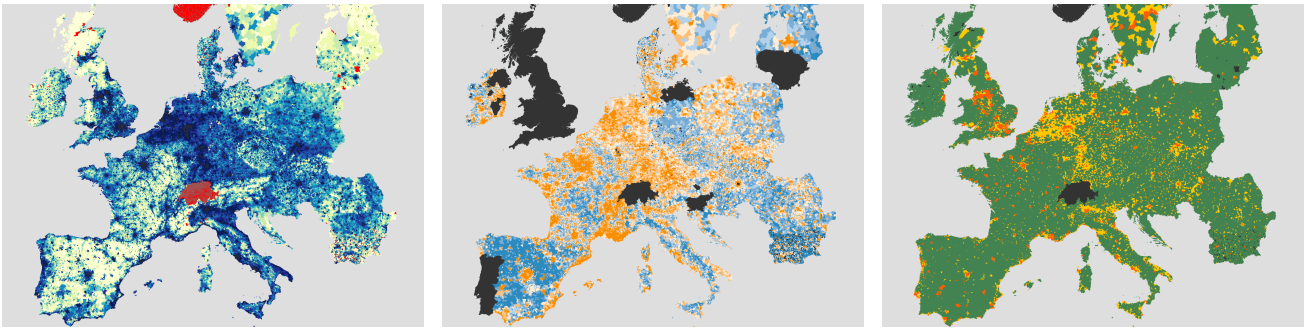


*Figure 1.: Population density (left), Population change 1961-2011 (middle), and Degree of urbanization (right) on LAU2 level*

### 4. Degree of urbanization

Eurostat administers each municipality's degree of urbanization on a scale of three, where one signifies that the municipality is highly urbanized, and three means a rural environment. Since this data was also available on a LAU2 level, we collected the dataset and intend to use it in analyzing the composition of the ESIF programs: regarding the amount of funding received, which countries were preferring rural areas to urban regions and vice versa?

### 5. Other datasets

Unfortunately, the number of publicly available datasets in LAU2 level is quite limited.

The European Statistical System does have a couple of municipality level datasets[4] that could be used to further enrich the analysis of the composition of EU funds, namely those which have been collected during the 2011 EU Census:

- Sex and age distribution of persons
- Type and size of families and households
- Type and occupancy statuses of living quarters and buildings

The European Observation Network, Territorial Development and Cohesion (ESPON) portal lists *GDP in LAU2 units*[5] among its datasets, but they did not respond to our requests for information so far.

Eurostat also has regional data, but the highest resolution is typically NUTS3 level. Still, to explore GDP, demographics could be a viable option, even if it is only available in aggregated form.

Another possibility to include further datasets is to harvest them programatically from the internet. Based on preparatory research, infrastructure type datasets could be easily collected. A couple of examples:

| METRIC FOR LAU2 | POSSIBLE SOURCE |
|---|---|
| ACCESS TO HOSPITALS (DISTANCE TO NEAREST IN KM$^2$) | http://hospitals.webometrics.info/en/europe |
| ACCESS TO UNIVERSITIES (DISTANCE TO NEAREST IN KM$^2$) | http://www.webometrics.info/en/Europe |
| ACCESS TO HIGHWAYS (DISTANCE TO NEAREST EXIT IN KM$^2$) | http://wiki.openstreetmap.org/wiki/Highways |

[4] Available at: https://ec.europa.eu/CensusHub2/query.do?step=selectHyperCube&qhc=false

[5] Available at: http://database.espon.eu/db2/resource?idCat=42

| METRIC FOR LAU2 | POSSIBLE SOURCE |
|---|---|
| ACCESS TO RAILWAY STATIONS (DISTANCE TO NEAREST IN KM$^2$) | http://wiki.openstreetmap.org/wiki/Railway_stations |
| ACCESS TO RESTAURANTS, SHOPS, STORES, GAS STATIONS, ETC. (NUMBER OF UNITS WITHIN RADIUS) | https://developers.google.com/places/ |

*Table 3.: Possible datasources to be harvested from the internet*

# Architecture

A major part of the research was building an architecture and creating a web application, where the results could be browsed. Designing the database and the final data schema of the project was a key achievement, since this is exactly what the individual, national and regional data sources publishing the subsidy transactions are lacking: a unified, standardized way of collecting and storing the data of the subsidies allocated to beneficiaries on a transactional level.

| DATABASE COLUMN | DESCRIPTION |
|---|---|
| transaction_id | An identifier for every transaction, unique across all countries. |
| country | Country reporting the transaction. |
| country_code | ISO Alpha-2 code for the country reporting the transaction. |
| fund_acronym | Acronym of fund, with following values: ERDF, ESF, CF. |
| funding_period | 2007-2013 or 2014-2020. |
| amount | Freely associated amount of subsidy in euros that is not the total of the project, neither paid by the EU. |
| amount_kind | Descriptive field for freely associated amount kind. |
| eu_cofinancing_amount | Subsidy paid by the EU in euros. |
| total_amount | Total cost of project in euros. |
| beneficiary_id | An identifier for every beneficiary, unique across all countries. |
| beneficiary_name | The name of the beneficiary. |
| beneficiary_country_code | ISO Alpha-2 code of the country where the beneficiary is registered. |
| beneficiary_country | The name of the country where the beneficiary is registered. |

| DATABASE COLUMN | DESCRIPTION |
|---|---|
| **beneficiary_state** | Name of the outmost administrative boundary unit where the beneficiary is registered. Usually equals to NUTS1. |
| **beneficiary_region** | Name of the administrative boundary unit where the beneficiary is registered. Usually equals to NUTS2. |
| **beneficiary_nuts3** | NUTS3 code of the boundary unit, where the beneficiary is registered. |
| **beneficiary_lau2** | LAU2 code of the municipality where the beneficiary is registered. |
| **beneficiary_city** | Name of the city where the beneficiary is registered. |
| **beneficiary_postal_code** | Postal code of the address, where the beneficiary is registered. |
| **beneficiary_address** | Postal address of the beneficiary. |
| **beneficiary_lat** | Latitude coordinate of the beneficiary's address, in WGS84 geodetic datum. |
| **benefciary_long** | Longitude coordinate of the beneficiary's address, in WGS84 geodetic datum. |
| **geocoding_state** | Boolean for geocoding state. |
| **project_name** | Description of the project. |
| **project_country** | Country name, where the project was realized. |
| **project_state** | Name of the outmost administrative boundary unit where the project was realized. Usually equals to NUTS1. |
| **project_region** | Name of the administrative boundary unit where where the project was realized. Usually equals to NUTS2. |
| **project_nuts3** | NUTS3 code of the boundary unit, where the project was realized. |
| **project_lau2** | LAU2 code of the boundary unit, where the project was realized. |
| **project_city** | Name of the city where the project was realized. |
| **project_postal_code** | Postal code of the address where the project was realized. |
| **project_address** | Postal address where the project was realized. |
| **project_lat** | Latitude coordinate of the project's address, in WGS84 geodetic datum. |

| DATABASE COLUMN | DESCRIPTION |
|---|---|
| **project_long** | Longitude coordinate of the project's address, in WGS84 geodetic datum. |

*Table 4.: The data model of the transaction table*

An important part of the project is to publish the results in a web based, interactive environment and make the collected data available for public use. Though the results are already available on the web, the following architecture is just a prototype, and will probably change significantly in the future. After scraping the data from a number of sources, the results are loaded using custom ETL scripts into an Amazon RDS for PostgreSQL database. The final results are stored in scale independent vector tile sets for the prototype - this will also likely be changed, mostly due to already hitting performance limitations. The final form of the application will probably feature the maps in raster format, and interactivity will be supported by UTFGrid. Map views will be complemented with interactive charts, and the result files will be available for downloading.

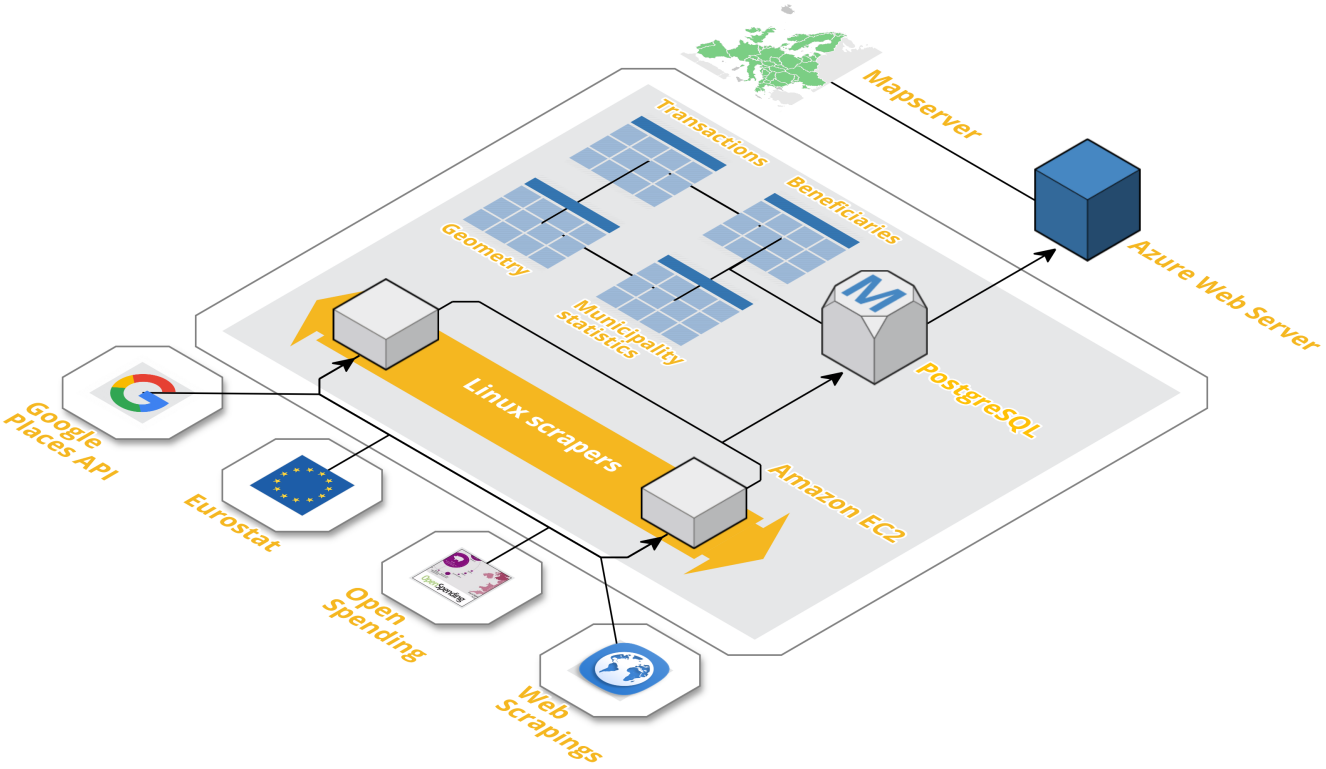



*Figure 2.: Current architecture of the web application*

# METHOD 1. for geolocating: using SubsidyStories as source and geocoding with the Google Places API

A prerequisite for our research was that the collection, cleaning and standardization of the transactional data is taken care of by SubsidyStories, and our work would be focused on attaining geolocation based on what is already available in the datasets collected and published by SubsidyStories[6].

The dataset was not always validated and we encountered many errors during the processing of the data - data currently published by the SubsidyStories project for France, the United Kingdom and Ireland is collected by us, and we helped them fix a number of errors in the Czech, Finnish and Hungarian datasets as well. Generally, when corrupted data was identified, we made sure to notify the SubsidyStories team, so at least on high level their dataset was harmonized with what we were working with locally. Comparing the amounts received by each country in the dataset with official figures published by the EU[7] we see that in some cases the sums show a magnitude of difference - these are continuously subject to reconsideration and are results of either interpretational differences (which funds or funding periods are included, what was the exchange rate used for converting non euro currencies, etc.) or missing data (at some edge cases whole programs are missing, at other places the funds paid by the member states and the funds paid by the EU are not separated and so on).

| COUNTRY | SUM OF ESIF 2007-2013 IN THE DATASET (€) | OFFICIAL FIGURES PUBLISHED BY EU (€) | RATIO (%) |
|---|---|---|---|
| POLAND | €76,058,575,974.90 | €67,185,549,244.00 | 113.20% |
| SPAIN | €49,926,158,500.50 | €34,657,733,981.00 | 144.05% |
| ITALY | €34,668,412,758.09 | €27,957,849,976.00 | 124.00% |
| CZECH REPUBLIC | €26,803,942,501.47 | €26,526,375,721.00 | 101.04% |
| GERMANY | €24,735,455,539.50 | €25,488,616,290.00 | 97.04% |
| HUNGARY | €38,456,152,525.35 | €24,921,148,600.00 | 154.31% |
| PORTUGAL | €23,656,615,168.82 | €21,411,560,512.00 | 110.48% |
| GREECE | €9,373,156,135.00 | €20,210,261,445.00 | 46.37% |
| FRANCE | €15,185,841,948.17 | €13,449,221,051.00 | 112.91% |
| SLOVAKIA | €12,295,960,193.92 | €11,498,331,484.00 | 106.93% |
| UNITED KINGDOM | €12,421,482,244.20 | €9,890,937,463.00 | 125.58% |

[6] Dataset is accessible here: http://subsidystories.eu/

[7] For validation this report was referred: https://cohesiondata.ec.europa.eu/dataset/Total-EU-Allocations-Per-MS-For-2007-2013/4taz-54g9

| COUNTRY | SUM OF ESIF 2007-2013 IN THE DATASET (€) | OFFICIAL FIGURES PUBLISHED BY EU (€) | RATIO (%) |
|---|---|---|---|
| LITHUANIA | €6,750,675,528.74 | €6,775,492,823.00 | 99.63% |
| BULGARIA | €19,485,852,923.60 | €6,673,628,244.00 | 291.98% |
| LATVIA | €1,137,691,776.00 | €4,530,447,634.00 | 25.11% |
| SLOVENIA | €4,576,151,396.24 | €4,101,048,636.00 | 111.58% |
| ESTONIA | €5,419,641,772.00 | €3,403,459,881.00 | 159.23% |
| BELGIUM | €233,688,915.42 | €2,063,500,766.00 | 11.32% |
| NETHERLANDS | €1,352,421,767.00 | €1,660,002,737.00 | 81.47% |
| SWEDEN | €1,078,214,010.18 | €1,626,091,888.00 | 66.30% |
| FINLAND | €2,857,841,905.00 | €1,595,966,044.00 | 179.06% |
| AUSTRIA | €2,513,015,220.69 | €1,204,478,581.00 | 208.63% |
| MALTA | €969,374,212.00 | €840,123,051.00 | 115.38% |
| IRELAND | €602,999,286.15 | €750,724,742.00 | 80.32% |
| CYPRUS | €741,960,746.40 | €612,434,992.00 | 121.14% |
| DENMARK | €491,807,388.90 | €509,577,239.00 | 96.51% |
| LUXEMBOURG | €30,852,529.89 | €50,487,332.00 | 61.10% |
| CROATIA | €461,660,027.15 | N/A | N/A |
| ROMANIA | N/A | N/A | N/A |
| **EU 26** | **€371,823,942,868.14** | **€319,595,050,357.00** | **116.34%** |

*Table 5.: Comparing country level sums of subsidies with official figures*

Since it was clear that this dataset did not contain any geolocations, the initial, and somewhat naive approach was to geocode the transactions by the name of the beneficiary, which was administered for each transaction, as required by the regulations of the European Commission. Because we had satisfactory results from former experiences with the Google Places API, we chose to use this service for attaining addresses, and eventually coordinates for each unique beneficiary. The SubsidyStories database holds 2,793,113 transactions, which are distributed among 1,059,573 unique beneficiaries (no natural language processing methods were applied to differentiate between different word forms of the same company names or other named entities). Geocoding over 1 million data-points was not trivial, but the initial results were quite satisfactory, as 71.44% of all beneficiaries could be associated with an address, resulting that 92.16% of all distinct transactions could be identified with a location and 97.70% of the transaction sums could be geocoded with this method.

| COUNTRY | NUMBER OF DISTINCT BENEFICIARIES | SUCCESFULLY GEOCODED BENEFICIARIES | SUCCESS RATIO (%) |
|---|---|---|---|
| SLOVENIA | 2,438 | 2,343 | 96.10% |
| LUXEMBOURG | 41 | 39 | 95.12% |
| DENMARK | 292 | 273 | 93.49% |
| FINLAND | 46 | 43 | 93.47% |
| CZECH REPUBLIC | 25,823 | 23,840 | 92.32% |
| SLOVAKIA | 5,061 | 4,623 | 91.34% |
| UNITED KINGDOM | 7,847 | 7,135 | 90.92% |
| LATVIA | 64 | 57 | 89.06% |
| NETHERLANDS | 1,085 | 966 | 89.03% |
| BELGIUM | 643 | 569 | 88.49% |
| LITHUANIA | 3,979 | 3,448 | 86.65% |
| AUSTRIA | 31,934 | 27,642 | 86.55% |
| SWEDEN | 865 | 728 | 84.16% |
| FRANCE | 26,253 | 21,831 | 83.15% |
| MALTA | 154 | 126 | 81.81% |
| GREECE | 176 | 142 | 80.68% |
| CROATIA | 899 | 706 | 78.53% |
| PORTUGAL | 23,231 | 18,150 | 78.12% |
| ITALY | 98,773 | 75,662 | 76.60% |
| POLAND | 54,369 | 41,325 | 76.00% |
| GERMANY | 157,660 | 118,421 | 75.11% |
| ESTONIA | 13,281 | 9,609 | 72.35% |
| BULGARIA | 6,480 | 4,460 | 68.82% |
| SPAIN | 541,437 | 362,068 | 66.87% |
| HUNGARY | 48,279 | 29,137 | 60.35% |
| CYPRUS | 1,018 | 437 | 42.92% |
| IRELAND | 7,445 | 3,193 | 42.88% |

| COUNTRY | NUMBER OF DISTINCT BENEFICIARIES | SUCCESFULLY GEOCODED BENEFICIARIES | SUCCESS RATIO (%) |
|---|---|---|---|
| EU 27 | 1,059,573 | 765,973 | 71.44% |

*Table 6.: Geocoding results based on beneficiary name serviced by Google API*

Unfortunately, after aggregating the results to LAU2 level, and mapping them it became clear that even though with the help of the Google API some 71.44% of all beneficiaries could be identified - though the confidence level of this method is unknown - the method to associate the location of the transaction with the identified address of the beneficiary is somewhat questionable in many cases. Upon examining the individual datasets it came to our attention that there is a reappearing pattern where a given beneficiary is funded a significant portion of the whole country's subsidies - beneficiaries receiving more than 1% of all national funds were present at virtually all EU member states, but sometimes this ratio reached as high as 10%. Clearly, these beneficiaries were - usually government owned - institutions themselves, who were only reallocating the money. For this single reason, distributing the transaction amounts purely on the basis of the identified beneficiaries' address became skewed: government institutions and agencies are usually located in the capitals of each country, or at least at regional centers. In other words, with this method the data did not distribute too well in many countries as apparent with France on *Figure 3*.

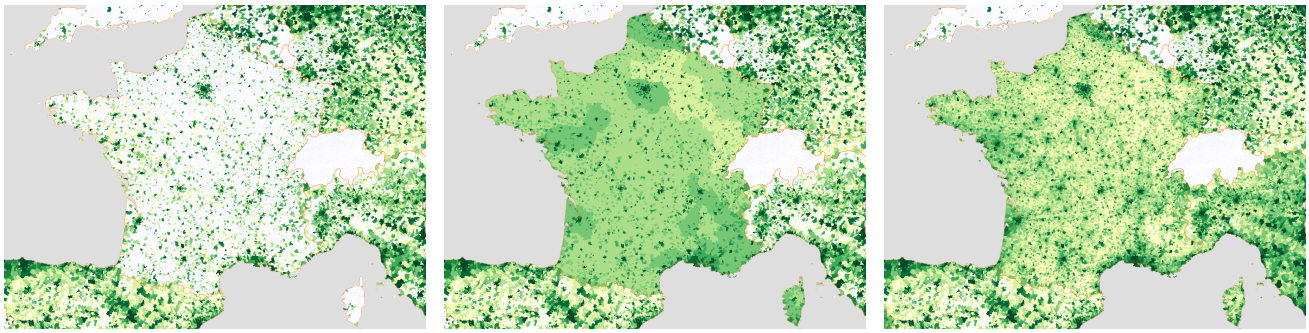


*Figure 3.: Distribution of French data based on Method 1. (left), Method 2a. (middle) and Method 2b. (right)*

Though the results could not be accepted as valid by any means, the distribution of transactions achieved by Method 1. did have its own analytical lessons. Firstly, there was no apparent correlation between the number of LAU2 units in each country, and the achieved distribution ratio - meaning that lower number of LAU2 units did not necessarily bring higher results. Secondly, the number of beneficiaries did not weigh in as much in influencing the results as one would expect. To support analytical evidence for these observations was out of the scope of this project - for now - as most of the results achieved with Method 1. were discarded later on.

| COUNTRY | NR. OF LAU2 WITH AT LEAST ONE TRANSACTION | NR. OF LAU2 WITH NO TRANSACTION | DISTRIBUTION RATIO (%) |
|---|---|---|---|
| ESTONIA | 223 | 3 | 98.67% |
| POLAND | 2,444 | 35 | 98.58% |
| SPAIN | 7,053 | 1,063 | 86.90% |
| CZECH REPUBLIC | 5,205 | 1,046 | 83.26% |
| AUSTRIA | 1897 | 460 | 80.48% |
| ITALY | 6,294 | 1,798 | 77.78% |
| HUNGARY | 2,164 | 990 | 68.61% |
| GERMANY | 7,555 | 3,858 | 66.19% |
| NETHERLANDS | 247 | 171 | 59.09% |
| SWEDEN | 169 | 121 | 58.27% |
| PORTUGAL | 2,271 | 1,989 | 53.30% |
| LITHUANIA | 277 | 263 | 51.29% |
| MALTA | 32 | 36 | 47.05% |
| SLOVAKIA | 1,358 | 1,569 | 46.39% |
| BELGIUM | 177 | 412 | 30.05% |
| CROATIA | 137 | 419 | 24.64% |
| IRELAND | 781 | 2,628 | 22.90% |
| UNITED KINGDOM | 2,080 | 8,223 | 20.18% |
| FRANCE | 6,075 | 30,603 | 16.56% |
| LATVIA | 13 | 106 | 10.92% |
| BULGARIA | 434 | 4,868 | 8.18% |
| DENMARK | 175 | 1,997 | 8.05% |
| CYPRUS | 43 | 572 | 6.99% |
| FINLAND | 20 | 316 | 5.95% |
| ROMANIA | N/A | N/A | N/A |
| LUXEMBOURG | N/A | N/A | N/A |
| SLOVENIA | N/A | N/A | N/A |
| GREECE | N/A | N/A | N/A |

| COUNTRY | NR. OF LAU2 WITH AT LEAST ONE TRANSACTION | NR. OF LAU2 WITH NO TRANSACTION | DISTRIBUTION RATIO (%) |
|---|---|---|---|
| EU 24 | 47,124 | 63,546 | 57.41% |

*Table 7.: Distribution of LAU2 units which could be paired with transaction data - Method 1.*

Still, the methodology to examine the geolocational distribution of subsidies within a given country on the basis of what percentage of receiving beneficiaries' address is identifiable could not be completely rejected: it does contain traces of transparency indicators (what percentage of the receiving ends has a publicly known address so to speak), though this methodology obviously has its throwbacks. As it later turned out, based on current data at some cases it is the only available programmatic methodology to geolocate a transaction - many times the name of the beneficiary is the only piece of information in the source data which could be used to trace the location of the beneficiary.

In overall, we decided to go back to source data, and look for traces of geolocations included.

## METHOD 2. for geolocating: reaching back to source files for geodata

After the failure of Method 1. there was no other way but to go back to source data as it is published by national and regional institutions, and scrape the datasets from their original form. The form of publication shows a great variance across sources: some countries offer machine readable files - though their data model is far from being harmonized or standardized, and are often corrupted to some extent - at other sources only .PDF files are accessible. Some other countries have online databases that could be queried - typically built in every imaginable way possible, so custom written scraper scripts were needed for each case.

| COUNTRY | DATA PORTAL ADDRESS |
|---|---|
| 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/rrtiepa/?lang=en |

| COUNTRY | DATA PORTAL ADDRESS |
|---|---|
| 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 | https://www.europaomdehoek.nl |
| POLAND | http://www.mapadotacji.gov.pl/en |
| PORTUGAL | https://www.portugal2020.pt/Portal2020 |
| ROMANIA | 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 |
| UNITED KINGDOM | https://www.gov.uk/government/publications/ESIF-useful-resources |

*Table 8.: Entry points to national ESIF data portals*[8]

Fortunately it turned out that source data often contains geolocational information about the transactions that was truncated from the SubsidyStories data. In many cases, the exact address, postal code or region (varying from NUTS2 to LAU2 level) of the beneficiary's headquarters was included in the source, but even better for our purposes, geolocational information about the realization of the project was frequently part of the original reports.

Project locations were typically published in an array like manner: for each transaction row, a list of project locations were given. The NUTS level and format of the administered project locations showed great variance: we encountered LAU2 level names and codes, LAU1

[8] Please note: in some cases (United Kingdom, Germany, Austria, etc.) subsidy data is published among a number of sources as administration is distributed among regional institutions.

names, NUTS3 names and codes, NUTS2 and NUTS1 names so far. At some places information suggests that the whole country benefited from the given transaction, describing the location of the project as “national”.

Since project locations even varied in level of aggregation within one row - meaning that for a given transaction, a list of regional locations equal to NUTS3 units and a list of LAU2 units could be provided - we decided to keep as much from the source data as possible, therefore we stuck to the lowest level of representation and tried to convert everything to LAU2 and then aggregate the data when necessary.

This also meant that a method for distribution had to be applied when the data was not given in LAU2 level. We experimented with distributing the sum of the transaction equally among LAU2 member units of the parent unit, but rejected this method (annotated as Method 2a. on *Figure 3.*) early on, exactly because of the equal nature of distribution. For example, if a large geolocation, like Île-de-France (which is a NUTS2 unit of France) was provided only, it seemed unfair to treat Paris and Orly (both same level child units of the given region) with the same weight, since the former has a hundred times more residents than the latter.

Eventually we decided to weight the distribution of transactions received with the population of each LAU2 unit: firstly, the per capita value of the transaction was calculated for the given region (the transaction sum divided by the population of the whole region that was provided), then each LAU2 member unit of the given location received amounts respective to their population (the per capita value of the whole region was multiplied by the population of the given LAU2 unit, and assigned to it).

This distribution method (annotated as Method 2b. on *Figure 3.*) was carried out on a transaction level, and resulted in exponentially increasing the number of rows in our database.

| COUNTRY | NR. OF LAU2 WITH AT LEAST ONE TRANSACTION | NR. OF LAU2 WITH NO TRANSACTION | RATIO (%) | REDISTRIBUTED? |
|---|---|---|---|---|
| FRANCE | 36,678 | 0 | 100% | Yes |
| ITALY | 8,092 | 0 | 100% | Yes |
| PORTUGAL | 4,260 | 0 | 100% | Yes |
| POLAND | 2,479 | 0 | 100% | Yes |
| CROATIA | 556 | 0 | 100% | Yes |
| FINLAND | 336 | 0 | 100% | Yes |
| SLOVENIA | 210 | 0 | 100% | No |
| CZECH REPUBLIC | 6,250 | 1 | 99.98% | Yes |
| SWEDEN | 289 | 1 | 99.65% | Yes |

| COUNTRY | NR. OF LAU2 WITH AT LEAST ONE TRANSACTION | NR. OF LAU2 WITH NO TRANSACTION | RATIO (%) | REDISTRIBUTED? |
|---|---|---|---|---|
| IRELAND | 3,374 | 35 | 98.97% | Yes |
| ESTONIA | 223 | 3 | 98.67% | No |
| SPAIN | 7,053 | 1,063 | 86.90% | No |
| AUSTRIA | 1,889 | 468 | 80.14% | No |
| HUNGARY | 2,164 | 990 | 68.61% | No |
| GERMANY | 7,555 | 3858 | 66.19% | No |
| NETHERLANDS | 247 | 171 | 59.09% | No |
| LITHUANIA | 277 | 263 | 51.29% | No |
| MALTA | 32 | 36 | 47.05% | No |
| SLOVAKIA | 1358 | 1569 | 46.39% | No |
| BELGIUM | 177 | 412 | 30.05% | No |
| UNITED KINGDOM | 2080 | 8223 | 20.18% | No |
| LATVIA | 13 | 106 | 10.92% | No |
| BULGARIA | 434 | 4868 | 8.18% | No |
| DENMARK | 175 | 1997 | 8.05% | No |
| CYPRUS | 43 | 572 | 6.99% | No |
| ROMANIA | N/A | N/A | N/A | No |
| GREECE | N/A | N/A | N/A | No |
| LUXEMBOURG | N/A | N/A | N/A | No |
| **EU 25** | **86,209** | **24,671** | **77.75%** | |

*Table 9.: Distribution of LAU2 units which could be paired with transaction data - Method 2b.*

So far, we could trace back and geocode information about the realization of the project for the following countries:

- Croatia
- Finland
- France
- Ireland
- Italy
- Poland
- Portugal

- Sweden

After re-scraping the datasets for these countries from source, and applying this method of distribution, the number of transactions for these countries grew from 1,237,838 rows to 88,688,957 rows, which is a magnitude of difference (+7164.82%), introducing new performance problems when working with the data.

| COUNTRY | NR. OF TRANS. BEFORE REDISTR. | NR. OF TRANS. AFTER REDISTR. | CHANGE (%) |
|---|---|---|---|
| FRANCE | 101,960 | 33,435,147 | +32,692.41% |
| ITALY | 929,709 | 30,284,091 | +3,157.37% |
| POLAND | 109,536 | 14,535,745 | +13,170.29% |
| PORTUGAL | 62,360 | 9,248,210 | +14,730.35% |
| IRELAND | 9,708 | 831,920 | +8,469.42% |
| SPAIN | 787,088 | 787,088 | 0 |
| GERMANY | 318,784 | 318,784 | 0 |
| FINLAND | 19,940 | 257,540 | +1,191.57% |
| HUNGARY | 126,083 | 126,083 | 0 |
| CZECH REPUBLIC | 124,216 | 124,216 | 0 |
| AUSTRIA | 76,578 | 76,578 | 0 |
| CROATIA | 2,532 | 65,056 | +2,469.35 |
| SLOVAKIA | 36,488 | 36,488 | 0 |
| SWEDEN | 2,123 | 31,248 | +1,371.87% |
| ESTONIA | 26,141 | 26,141 | 0 |
| UNITED KINGDOM | 16,786 | 16,786 | 0 |
| BULGARIA | 11,798 | 11,798 | 0 |
| LITHUANIA | 10,778 | 10,778 | 0 |
| SLOVENIA | 5,234 | 5,234 | 0 |
| GREECE | 2,079 | 2,079 | 0 |
| NETHERLANDS | 1,808 | 1,808 | 0 |
| CYPRUS | 1,474 | 1,474 | 0 |
| BELGIUM | 1,166 | 1,166 | 0 |

| COUNTRY | NR. OF TRANS. BEFORE REDISTR. | NR. OF TRANS. AFTER REDISTR. | CHANGE (%) |
|---|---|---|---|
| DENMARK | 705 | 705 | 0 |
| MALTA | 273 | 273 | 0 |
| LATVIA | 134 | 134 | 0 |
| LUXEMBURG | 75 | 75 | 0 |
| ROMANIA | N/A | N/A | N/A |
| EU 27 | 2,785,556 | 90,236,645 | +3,139.44% |

*Table 10.: Number of transactions before and after redistribution per country*

One could argue the fairness and validity of the distribution of EU subsidies based on weighting with population data, since this applies that the subsidies are accessible for everyone regardless of their place of residence within a given territorial unit, which is most likely not the case. Still, we feel confident that this methodology makes the data comparable on higher aggregation levels, and skewing is negligible when looking at national levels. We are quite certain that there is added value to processing the datasets with this methodology, even though it won't be applicable for every member of the EU 28, since not all datasets include geolocational information. For these countries, we intend to include geocoding based on Method 1. (looking up the beneficiaries' address), and introduce a way of weighting to filter out beneficiaries, who received a significant amount from all subsidies allocated to the country. For these transactions, we intend to introduce distribution across the whole country.

## Published results and further steps

The results of our research are already publicly accessible, though it is clearly a work in progress. After finishing the re-scraping of all EU 28 member states and redistributing all published transactions, we intend to share the data with the public in machine readable format, as well as in an interactive, browsable way.

| VIEW | ACCESSIBLE AT |
|---|---|
| POPULATION DENSITY - 2011 (CAPITA / KM$^2$) | http://subsidies.westeurope.cloudapp.azure.com/styles/flat/#3.3/52.31/10.76 |
| POPULATION CHANGE - 1961 - 2011 (%) | http://subsidies.westeurope.cloudapp.azure.com/styles/eupop3/#3.3/52.31/10.76 |
| DEGREE OF URBANIZATION | http://subsidies.westeurope.cloudapp.azure.com/styles/urban/#3.3/52.31/10.76 |
| SUBSIDIES RECEIVED | http://subsidies.westeurope.cloudapp.azure.com/styles/lautrans/#3.3/52.31/10.76 |

| VIEW | ACCESSIBLE AT |
|---|---|
| SUBSIDIES RECEIVED PER CAPITA | http://subsidies.westeurope.cloudapp.azure.com/styles/lautranscap/#3.3/52.31/10.76 |

*Table 11.: Access to published results*

Geolocating the data is only the first step in understanding the composition and grant effectiveness of the ESIF 2007 - 2013 program. We are excited about the possibilities that geolocated data could present in this space - here's just a brief list of analytical perspectives that are almost within reach:

- How much funding was received per capita in each municipality? Are there any recognizable geopolitical trends, are they inline with the EU's cohesion policies?
- Is the EU supporting the younger or the older population (correlations with age distribution in each municipality)? How about migration trends inside the EU? Communities with decreasing or increasing populations are funded better? What about urban / rural contrasts?
- Does the advancement of the local infrastructure (access to hospitals, transportation, education, commercial services, etc.) influence in any way the success of receiving funds? Are there local, regional, national or EU-wide trends in this?

We believe that making these observations public - among with the data model as a reference for future subsidy programs within the EU - is an effective way of convincing policy makers and the EU bureaucracy to take further steps in providing transparency not just for how the EU funds were spent, but also on what their impact was.

# Sources

Council Regulation (EC) No 1083/2006 of 11 July 2006
http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32006R1083

Council Regulation (EC) No 1828/2006 of 8 December 2006
http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32006R1828

LAU2 geometries for the EU 28
http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/communes

Historical population data for LAU2 units for the EU 28
http://ec.europa.eu/eurostat/web/nuts/local-administrative-units

Degree of urbanization for LAU2 units for the the EU 28
http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/degurba

ESIF allocations per member state for the 2007 - 2013 program
https://cohesiondata.ec.europa.eu/dataset/Total-EU-Allocations-Per-MS-For-2007-2013/4taz-54g9

Regional datasets offered by Eurostat
http://ec.europa.eu/eurostat/web/regions/data/database

Local data offered by ESPON
http://database.espon.eu/db2/resource?idCat=42

EU Census Hub data
https://ec.europa.eu/CensusHub2/query.do?step=selectHyperCube&qhc=false

SubsidyStories website
http://subsidystories.eu/