A pilot project to collect, clean and assess the list of operations produced by cohesion policy programmes at national, regional level

Final Report

2020CE16BAT015

Balazs Krich
August - 2020

Regional and Urban Policy

Disclaimer

The information and views set out in this report are those of the author and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

EUROPEAN COMMISSION

Directorate-General for Regional and Urban Policy
Directorate B — Policy
Unit B.2 Evaluation and European Semester

Contact: John WALSH

E-mail: REGIO-EVAL@ec.europa.eu

European Commission
B-1049 Brussels

A pilot project to collect, clean and asses the list of operations produced by cohesion policy programmes at national, regional level

Final Report

Directorate-General for Regional and Urban Policy

2020 EN

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Manuscript completed in September 2020

Luxembourg: Publications Office of the European Union, 2020

ISBN 978-92-76-23871-3
doi: 10.2776/82501

TABLE OF CONTENTS

ABSTRACT

The goal of the Pilot project to collect, clean and asses the list of operations produced by cohesion policy programmes at national, regional level was to create a database for the operations in the 2014 - 2020 funding cycle among the EU 27 member states and the United Kingdom, including Interreg programmes. The project also delivered a software solution that self-documents the collection of named datasets, and makes all operations and procedures (accessing, transforming, cleaning, converting, etc.) programmatically reproducible, automated and extendible.

The initial list of data sources was provided by the European Commission. It was not within the scope of the project to research and collect all transactions within the 2014 - 2020 funding cycle - upon agreement only file-based, programmatically accessible and computer readable sources were included, therefore data only existing in HTML format for example was outside of the scope. As the presented framework is extendible, it allows the technical user to include additional data sources that were not part of the project originally.

Upon completing the project, the master database included data from 22 member states (plus Interreg programmes), 716 693 transactions from 609 940 operations, in the total value of € 288 458 996 260.21.

EXECUTIVE SUMMARY

Collecting data from national and regional sites is a difficult challenge because the formats and standards of data storage and presentation show extremely high variance. This becomes even more challenging when considering the quality of the accessible files: in most cases the creators of the content are clearly not information technology professionals. What might pass as an internal report or a file for accounting / presentational purposes dedicated for the human end user, will fail for many reasons when we try to access it programmatically and through programming scripts. However, there are clearly recognisable patterns in the presentation of named datasets, even across multiple countries, languages and cultures within the European Union. By applying predefined rulesets to these recognised patterns, it becomes possible to collect, standardise and normalise the data.

This pilot project was able to collect data from 22 member states (for Cyprus, Estonia, Finland, Lithuania, Malta and Sweden the project was unable to detect programmatically accessible, file-based data sources) and 37 Interreg programmes, which all together presented 167 independent sources. The final dataset contains 716 693 transactions from 609 940 operations, in the total estimated value of € 288 458 996 260.21.

After defining the exact location of the accessible files themselves and making the selections for further processing, it was agreed that only file-based data sources will be included. The formats themselves were mostly Microsoft Excel files, occasionally .csv or XML files. Upon request the presented software solution included the possibility to manually include files which were not available at a given web address - the Hungarian dataset was collected as such.

Another challenge was associating a direct, one-to-one relationship between the files' contents and the CCI programme codes provided by the representatives of the Commission. For 128 CCI programmes an unambiguous link was detectable - a further development could be to look for traces of the codes in the raw data and try to extract further links.

Upon first inspection, the files might seem tabular in structure for the human user, but they prove to be invalid when trying to load them into a database. A significant effort of the project was to extract valid tabular structures from the source files: this included detecting a valid header and a fixed number of columns across all rows of the files. Different language settings and character encodings, compressions, merged cells, corrupted files and multiple tabs within one file - but not necessarily containing tabular data - were among the key difficulties, which the project overcame. The outcome of these steps are presented in detail in Appendix A, the applied methodologies are described in Appendix B - both being inseparable parts of this report.

Once the tabular structure was ensured - while maintaining the original column order, and a connection between the source files and the database tables, so the source data and the output remained comparable - it became possible to carry out transformations on the data in a self-documenting way via SQL scripts.

Initially an automation was considered for this task, but the data sources and the formats showed such a high variance that the idea of a general solution was discarded. Instead, transformations were carried out in the scope of a single table, which had two major benefits. Firstly, the possibility of errors and incorrect transformations was reduced significantly as every table's every column was treated individually with human inspection. But possibly even more importantly, now the transformation rules are changeable, expandable, modifiable independently in the scope of a single table, allowing space for corrections and improvements.

We believe that the output of this self-documenting approach - the close to 10 000 lines of SQL scripts - presents the real value of the project: there is no single manual, undocumented step in the entire scope of the project, therefore testing, bug-fixing, change requests are transparent and exposed as it would be the expectation from a software project.

The quality of the data was surprisingly high when compared to the quality of the original sources: from the 11 compulsory fields to be provided, 9 were extracted and transformed into valid data types over 95% of the cases, with no missing values.

The detection of geolocational data also proved to be promising: about 98.7% of all rows contained some type of geolocational data. The project included an initial, naive approach to categorise and distribute the detected geolocational data among the NUTS / LAU statistical system. Though the extraction of location names, validation of detected codes and the pairing of them with the location names was outside of the scope of this project, upon the initial results we suspect that a high matching rate could be achieved.

The project also provided a mapping for the additional, non compulsory fields in the source files and standardised those into a fixed structure. These proved to be rather fuzzy, containing a lot of ambiguity. A clear definition of expected values and possible patterns in detecting them could be a further improvement, which could be a valuable extension of the current codebase.

FINAL DATA STRUCTURE OF THE MASTER DATABASE

The final data structure of the master database is designed as presented in the following table.

Column name	Data type	Description
operation_name	TEXT	Name of the project / operation.
beneficiary_name	TEXT	Name of the lead beneficiary.
operation_summary	TEXT	Short description of the project.
operation_start_date	DATE	Date format is YYYY-MM-DD.
operation_end_date	DATE	Date format is YYYY-MM-DD.
currency	TEXT	Native detected currency.
operation_total_expenditure	NUMERIC	In native currency.
eu_cofinancing_rate	NUMERIC	Presented in numeric values between 0 and 100.
country	TEXT	If a country column was provided, values are mapped here. No value consolidation is included.
operation_location	TEXT	If a geolocational data type was provided, it is included here. If multiple separate fields are included, always the column containing the highest precision is included - practically postal codes.
code_of_category_intervention	TEXT	If an identifier is included, it is mapped here.
name_of_category_intervention	TEXT	If a name or a name and an identifier included, it is mapped here.
date_of_last_update	DATE	Date format is YYYY-MM-DD.
operation_nuts0	TEXT	Alpha-2 codes of the country according to the ISO 3166 standard. Always filled, even if country column contains no values.
operation_nuts1_code	TEXT	Non validated / suspected NUTS1 code of geolocation.
operation_nuts1_name	TEXT	Non validated / suspected NUTS1 name of geolocation.
operation_nuts2_code	TEXT	Non validated / suspected NUTS2 code of geolocation.
operation_nuts2_name	TEXT	Non validated / suspected NUTS2 name of geolocation.
operation_nuts3_code	TEXT	Non validated / suspected NUTS3 code of geolocation.
operation_nuts3_name	TEXT	Non validated / suspected NUTS3 name of geolocation.
operation_lau1_code	TEXT	Non validated / suspected LAU1 code of geolocation.
operation_lau1_name	TEXT	Non validated / suspected LAU1 name of geolocation.
operation_lau2_code	TEXT	Non validated / suspected LAU2 code of geolocation.
operation_lau2_name	TEXT	Non validated / suspected LAU2 name of geolocation.
operation_city	TEXT	If geolocational data could not be suspected, values are included here.
operation_district	TEXT	If district like information is provided, it is included here.
operation_zip_code	TEXT	Non validated postal codes.
member_state_value	NUMERIC	If member state contribution is included separately, it is mapped here. In native currency.
eu_subsidy_value	NUMERIC	If EU contribution is included separately, it is included here. In native currency.
beneficiary_id	TEXT	If a beneficiary identifier is included, it is mapped here.
operation_id	TEXT	If an operation identifier is provided, it is mapped here.

priority_axis	TEXT	If any information about priority axis is provided, it is included here.
form_of_finance	TEXT	If any information about form of finance is provided, it is included here.
territorial_dimension	TEXT	If any information about territorial dimension is provided, it is included here.
territorial_delivery_mechanism	TEXT	If any information about territorial delivery mechanism is provided, it is included here.
esf_secondary_theme	TEXT	If any information about ESF secondary theme is provided, it is included here.
economic_dimension	TEXT	If any information about economic dimension is provided, it is included here.
cci	TEXT	CCI programme code, if identifiable. Otherwise truncated first 6 digits of CCI code. Always filled.
data_source	TEXT	Unique internal identifier of the source table after transformations, helping to trace back the original source of the data. Always filled.
rate	NUMERIC	2019 yealy average exchange rate provided by Eurostat.
operation_total_expenditure_eur	NUMERIC	Exchanged to EUR using fixed rate column.
eu_subsdidy_value_eur	NUMERIC	Exchanged to EUR using fixed rate column.

Figure 1.: data structure of the master data base table

PRESENCE OR ABSENCE OF COMPULSORY COLUMNS AT THE WHOLE DATASET

Generally, the 11 compulsory columns are almost always present at each national and regional dataset, with two exceptions **Country** where only 32.5% of transactions have a value assigned and **Date of last update** where only 61.06 % of the rows have a value.

Please note: some of the missing value could be due to invalid or ambiguous formats (non existent date or ambiguous numerical value).

Column name	Transaction count	Present	Percentage (%)
Operation name	716,693	716,669	100
Beneficiary name	716,693	715,640	99.85
Operation summary	716,693	682,901	95.29
Operation start date	716,693	700,273	97.71
Operation end date	716,693	709,431	98.99
Operation total expenditure	716,693	716,133	99.92
EU cofinancing rate	716,693	713,944	99.62
Country	716,693	232,940	32.5
Operation location	716,693	707,490	98.72
Name of category intervention	716,693	704,707	98.33
Date of last update	716,693	437,607	61.06

Figure 2.: presence of compulsory columns at the whole dataset

PRESENCE OR ABSENCE OF COMPULSORY COLUMNS PER NATIONAL DATASET

The following table shows how the presence of compulsory columns distributes among national datasets, making it easy to identify which countries contain the biggest gaps.

Country	Column name	Transaction count	Present	Percentage (%)
AT	Operation name	1,264	1,264	100
AT	Beneficiary name	1,264	1,264	100
AT	Operation summary	1,264	1,264	100
AT	Operation start date	1,264	1,264	100
AT	Operation end date	1,264	1,264	100
AT	Operation total expenditure	1,264	1,264	100
AT	EU cofinancing rate	1,264	1,264	100
AT	Country	1,264	1,264	100
AT	Operation location	1,264	1,264	100
AT	Name of category intervention	1,264	1,264	100
AT	Date of last update	1,264	0	0
BE	Operation name	3,959	3,959	100
BE	Beneficiary name	3,959	3,958	99.97
BE	Operation summary	3,959	3,959	100
BE	Operation start date	3,959	3,959	100
BE	Operation end date	3,959	3,959	100
BE	Operation total expenditure	3,959	3,959	100
BE	EU cofinancing rate	3,959	3,959	100
BE	Country	3,959	3,756	94.87
BE	Operation location	3,959	3,955	99.9
BE	Name of category intervention	3,959	3,957	99.95
BE	Date of last update	3,959	0	0
BG	Operation name	2,634	2,634	100
BG	Beneficiary name	2,634	2,634	100
BG	Operation summary	2,634	2,634	100
BG	Operation start date	2,634	2,634	100
BG	Operation end date	2,634	2,634	100
BG	Operation total expenditure	2,634	2,634	100
BG	EU cofinancing rate	2,634	2,634	100
BG	Country	2,634	0	0
BG	Operation location	2,634	2,634	100
BG	Name of category intervention	2,634	2,634	100
BG	Date of last update	2,634	2,634	100
CZ	Operation name	33,477	33,477	100
CZ	Beneficiary name	33,477	33,477	100
CZ	Operation summary	33,477	33,470	99.98
CZ	Operation start date	33,477	24,580	73.42
CZ	Operation end date	33,477	33,477	100
CZ	Operation total expenditure	33,477	33,477	100
CZ	EU cofinancing rate	33,477	33,477	100
CZ	Country	33,477	33,477	100
CZ	Operation location	33,477	33,474	99.99
CZ	Name of category intervention	33,477	33,477	100
CZ	Date of last update	33,477	0	0
DE	Operation name	41,670	41,663	99.98
DE	Beneficiary name	41,670	41,670	100
DE	Operation summary	41,670	41,626	99.89

DE	Operation start date	41,670	41,200	98.87
DE	Operation end date	41,670	41,199	98.87
DE	Operation total expenditure	41,670	41,663	99.98
DE	EU cofinancing rate	41,670	41,200	98.87
DE	Country	41,670	40,390	96.93
DE	Operation location	41,670	41,567	99.75
DE	Name of category intervention	41,670	41,200	98.87
DE	Date of last update	41,670	0	0
DK	Operation name	287	287	100
DK	Beneficiary name	287	287	100
DK	Operation summary	287	285	99.3
DK	Operation start date	287	287	100
DK	Operation end date	287	287	100
DK	Operation total expenditure	287	287	100
DK	EU cofinancing rate	287	287	100
DK	Country	287	287	100
DK	Operation location	287	287	100
DK	Name of category intervention	287	287	100
DK	Date of last update	287	0	0
ES	Operation name	61,171	61,171	100
ES	Beneficiary name	61,171	61,171	100
ES	Operation summary	61,171	61,170	100
ES	Operation start date	61,171	61,171	100
ES	Operation end date	61,171	61,167	99.99
ES	Operation total expenditure	61,171	61,171	100
ES	EU cofinancing rate	61,171	61,171	100
ES	Country	61,171	61,171	100
ES	Operation location	61,171	61,001	99.72
ES	Name of category intervention	61,171	61,154	99.97
ES	Date of last update	61,171	0	0
FR	Operation name	7,490	7,490	100
FR	Beneficiary name	7,490	7,490	100
FR	Operation summary	7,490	6,455	86.18
FR	Operation start date	7,490	6,931	92.54
FR	Operation end date	7,490	7,264	96.98
FR	Operation total expenditure	7,490	7,473	99.77
FR	EU cofinancing rate	7,490	7,490	100
FR	Country	7,490	4,446	59.36
FR	Operation location	7,490	7,458	99.57
FR	Name of category intervention	7,490	6,899	92.11
FR	Date of last update	7,490	260	3.47
GR	Operation name	23,570	23,570	100
GR	Beneficiary name	23,570	23,570	100
GR	Operation summary	23,570	23,470	99.58
GR	Operation start date	23,570	23,570	100
GR	Operation end date	23,570	23,570	100
GR	Operation total expenditure	23,570	23,570	100
GR	EU cofinancing rate	23,570	23,570	100
GR	Country	23,570	23,470	99.58
GR	Operation location	23,570	23,470	99.58
GR	Name of category intervention	23,570	23,570	100
GR	Date of last update	23,570	23,470	99.58
HR	Operation name	2,666	2,666	100
HR	Beneficiary name	2,666	2,666	100
HR	Operation summary	2,666	2,666	100
HR	Operation start date	2,666	2,666	100
HR	Operation end date	2,666	2,666	100

HR	Operation total expenditure	2,666	2,666	100
HR	EU cofinancing rate	2,666	2,666	100
HR	Country	2,666	0	0
HR	Operation location	2,666	2,666	100
HR	Name of category intervention	2,666	2,666	100
HR	Date of last update	2,666	0	0
HU	Operation name	37,557	37,557	100
HU	Beneficiary name	37,557	37,557	100
HU	Operation summary	37,557	37,557	100
HU	Operation start date	37,557	37,542	99.96
HU	Operation end date	37,557	37,542	99.96
HU	Operation total expenditure	37,557	37,557	100
HU	EU cofinancing rate	37,557	37,414	99.62
HU	Country	37,557	37,557	100
HU	Operation location	37,557	37,309	99.34
HU	Name of category intervention	37,557	33,292	88.64
HU	Date of last update	37,557	0	0
IE	Operation name	772	772	100
IE	Beneficiary name	772	771	99.87
IE	Operation summary	772	772	100
IE	Operation start date	772	772	100
IE	Operation end date	772	700	90.67
IE	Operation total expenditure	772	772	100
IE	EU cofinancing rate	772	772	100
IE	Country	772	772	100
IE	Operation location	772	772	100
IE	Name of category intervention	772	352	45.6
IE	Date of last update	772	0	0
IT	Operation name	391,497	391,497	100
IT	Beneficiary name	391,497	391,497	100
IT	Operation summary	391,497	391,497	100
IT	Operation start date	391,497	391,497	100
IT	Operation end date	391,497	391,497	100
IT	Operation total expenditure	391,497	391,497	100
IT	EU cofinancing rate	391,497	389,823	99.57
IT	Country	391,497	0	0
IT	Operation location	391,497	391,497	100
IT	Name of category intervention	391,497	391,497	100
IT	Date of last update	391,497	391,497	100
LU	Operation name	22	22	100
LU	Beneficiary name	22	22	100
LU	Operation summary	22	22	100
LU	Operation start date	22	22	100
LU	Operation end date	22	22	100
LU	Operation total expenditure	22	22	100
LU	EU cofinancing rate	22	22	100
LU	Country	22	0	0
LU	Operation location	22	22	100
LU	Name of category intervention	22	22	100
LU	Date of last update	22	0	0
LV	Operation name	1,774	1,774	100
LV	Beneficiary name	1,774	1,774	100
LV	Operation summary	1,774	1,774	100
LV	Operation start date	1,774	1,774	100
LV	Operation end date	1,774	1,774	100
LV	Operation total expenditure	1,774	1,774	100
LV	EU cofinancing rate	1,774	1,774	100
LV	Country	1,774	0	0

LV	Operation location	1,774	1,774	100
LV	Name of category intervention	1,774	1,774	100
LV	Date of last update	1,774	1,774	100
NL	Operation name	662	649	98.04
NL	Beneficiary name	662	662	100
NL	Operation summary	662	658	99.4
NL	Operation start date	662	652	98.49
NL	Operation end date	662	662	100
NL	Operation total expenditure	662	662	100
NL	EU cofinancing rate	662	662	100
NL	Country	662	661	99.85
NL	Operation location	662	661	99.85
NL	Name of category intervention	662	509	76.89
NL	Date of last update	662	149	22.51
PL	Operation name	38,065	38,065	100
PL	Beneficiary name	38,065	38,063	99.99
PL	Operation summary	38,065	34,155	89.73
PL	Operation start date	38,065	38,057	99.98
PL	Operation end date	38,065	38,057	99.98
PL	Operation total expenditure	38,065	38,061	99.99
PL	EU cofinancing rate	38,065	38,053	99.97
PL	Country	38,065	211	0.55
PL	Operation location	38,065	36,381	95.58
PL	Name of category intervention	38,065	37,692	99.02
PL	Date of last update	38,065	0	0
PT	Operation name	37,527	37,527	100
PT	Beneficiary name	37,527	36,557	97.42
PT	Operation summary	37,527	13,072	34.83
PT	Operation start date	37,527	37,208	99.15
PT	Operation end date	37,527	37,209	99.15
PT	Operation total expenditure	37,527	37,527	100
PT	EU cofinancing rate	37,527	37,527	100
PT	Country	37,527	8,695	23.17
PT	Operation location	37,527	33,400	89
PT	Name of category intervention	37,527	37,524	99.99
PT	Date of last update	37,527	1,894	5.05
RO	Operation name	2,261	2,261	100
RO	Beneficiary name	2,261	2,261	100
RO	Operation summary	2,261	626	27.69
RO	Operation start date	2,261	621	27.47
RO	Operation end date	2,261	617	27.29
RO	Operation total expenditure	2,261	2,261	100
RO	EU cofinancing rate	2,261	2,261	100
RO	Country	2,261	0	0
RO	Operation location	2,261	1,884	83.33
RO	Name of category intervention	2,261	893	39.5
RO	Date of last update	2,261	0	0
SI	Operation name	4,944	4,944	100
SI	Beneficiary name	4,944	4,944	100
SI	Operation summary	4,944	4,944	100
SI	Operation start date	4,944	4,944	100
SI	Operation end date	4,944	4,944	100
SI	Operation total expenditure	4,944	4,893	98.97
SI	EU cofinancing rate	4,944	4,707	95.21
SI	Country	4,944	4,941	99.94
SI	Operation location	4,944	4,944	100
SI	Name of category intervention	4,944	4,941	99.94

SI	Date of last update	4,944	4,941	99.94
SK	Operation name	11,273	11,273	100
SK	Beneficiary name	11,273	11,199	99.34
SK	Operation summary	11,273	11,198	99.33
SK	Operation start date	11,273	9,175	81.39
SK	Operation end date	11,273	9,175	81.39
SK	Operation total expenditure	11,273	11,198	99.33
SK	EU cofinancing rate	11,273	11,198	99.33
SK	Country	11,273	568	5.04
SK	Operation location	11,273	9,100	80.72
SK	Name of category intervention	11,273	8,962	79.5
SK	Date of last update	11,273	8,607	76.35
TC	Operation name	10,357	10,353	99.96
TC	Beneficiary name	10,357	10,352	99.95
TC	Operation summary	10,357	7,906	76.33
TC	Operation start date	10,357	7,986	77.11
TC	Operation end date	10,357	7,982	77.07
TC	Operation total expenditure	10,357	9,952	96.09
TC	EU cofinancing rate	10,357	10,219	98.67
TC	Country	10,357	10,027	96.81
TC	Operation location	10,357	10,249	98.96
TC	Name of category intervention	10,357	8,348	80.6
TC	Date of last update	10,357	2,381	22.99
UK	Operation name	1,794	1,794	100
UK	Beneficiary name	1,794	1,794	100
UK	Operation summary	1,794	1,721	95.93
UK	Operation start date	1,794	1,761	98.16
UK	Operation end date	1,794	1,763	98.27
UK	Operation total expenditure	1,794	1,793	99.94
UK	EU cofinancing rate	1,794	1,794	100
UK	Country	1,794	1,247	69.51
UK	Operation location	1,794	1,721	95.93
UK	Name of category intervention	1,794	1,793	99.94
UK	Date of last update	1,794	0	0

Figure 3.: presence of compulsory columns per national dataset

NUMBER OF TRANSACTIONS, NUMBER OF BENEFICIARIES, NUMBER OF OPERATIONS, FINANCIAL DATA AT THE WHOLE DATASET

The following table contains the general summary of transactions, operations, beneficiaries and basic financial data of the whole dataset. Subsidies allocated is calculated row by row using the following formula:

*Subsidies allocated = Operation total expenditure * (EU cofinancing rate /100)*

Please note: a distinct operation name could appear multiple times, hence the difference between count of operations and transactions.

Transaction count	Operation count	Beneficiary count	Operation total expenditure (€)	Subsidies allocated (€)	AVG. EU co-financing rate (%)
716,693	609,940	218,705	288,458,996,260	190,221,945,336	64.96

Figure 4.: general summary of the whole dataset

NUMBER OF TRANSACTIONS, NUMBER OF BENEFICIARIES, NUMBER OF OPERATIONS, FINANCIAL DATA PER NATIONAL DATASET

The general summary of transactions, operations, beneficiaries and basic financial data shows the following distribution per country:

Country	Trans. count	Op. count	Benef. count	Operation total expenditure (€)	Subsidies allocated (€)	AVG. EU cofinancing rate (%)
AT	1,264	801	1,009	2,181,648,943	550,906,131	30.52
BE	3,959	3,542	1,805	4,997,300,999	2,069,540,996	30.65
BG	2,634	2,612	2,213	1,946,219,609	1,654,286,335	84.77
CZ	33,477	32,491	19,449	20,163,661,064	14,200,393,982	72.01
DE	41,670	26,091	17,029	14,615,867,607	7,904,267,169	66.52
DK	287	273	122	741,311,147	351,224,893	47.86
ES	61,171	61,171	27,106	20,502,053,308	14,022,758,124	67.84
FR	7,490	6,884	3,740	6,641,899,998	2,189,987,120	35.2
GR	23,570	22,687	17,787	4,603,913,335	3,578,213,096	76.99
HR	2,666	1,295	1,043	1,591,079,224	1,346,489,505	88.01
HU	37,557	37,254	20,498	26,366,511,290	19,760,982,256	70.62
IE	772	709	272	177,590,352	88,795,176	50
IT	391,497	343,335	60,074	42,230,578,075	23,314,494,686	62.12
LU	22	22	13	45,027,038	18,010,815	40
LV	1,774	1,768	703	5,081,615,815	3,345,932,062	67.92
NL	662	649	586	1,168,261,101	384,324,635	34.48
PL	38,065	31,744	17,062	42,210,298,080	30,115,907,515	76.6
PT	37,527	15,164	9,526	21,217,348,956	12,637,838,936	57.99
RO	2,261	1,999	1,901	35,172,191,617	27,390,810,045	67.1
SI	4,944	4,262	2,726	3,711,232,458	3,089,348,371	80.18
SK	11,273	10,878	6,034	13,975,247,921	11,156,285,210	82.38
TC	10,357	2,759	7,697	5,332,910,776	3,561,202,916	67.11
UK	1,794	1,738	673	13,785,227,551	7,489,945,362	53.56

Figure 5.: general summary per national dataset

NUMBER OF TRANSACTIONS, NUMBER OF BENEFICIARIES, NUMBER OF OPERATIONS, FINANCIAL DATA PER CCI PROGRAMME

The summary distribution among CCI programmes is included in the following table.

Please note: at certain case it was not possible to assign a CCI code to a dataset. In such cases the CCI code was truncated to the first 6 digits and multiple CCI programmes were merged under that identifier.

Programme title	Trans. count	Op. count	Ben. count	Operation total expenditure (€)	Subsidies allocated (€)
Investments in Growth and Employment - AT - ERDF	1,264	801	1,009	2,181,648,943	550,906,131
Brussels Capital Region - ERDF	3,036	2,951	1,759	3,143,754,888	1,314,272,591
Flanders - ERDF	203	202	148	405,805,547	149,232,164
Wallonia - ERDF	720	718	292	1,447,740,564	606,036,241
Science and Education for Smart Growth - BG - ESF/ERDF	159	159	146	149,692,409	127,238,548

Transport and transport infrastructure - BG - ERDF/CF	17	17	9	46,430,961	39,466,317
Environment - BG - ERDF/CF	30	30	25	69,959,996	59,465,997
Regions in Growth - BG - ERDF	416	413	164	668,609,853	568,318,043
Innovations and Competitiveness - BG - ERDF	2,012	1,993	1,888	1,011,526,390	859,797,431
2014CZ (CCI code / programme title unretrievable)	33,477	32,491	19,449	20,163,661,064	14,200,393,982
Niedersachsen - ERDF/ESF	11,509	4,040	3,372	461,152,927	244,673,507
Bayern - ERDF	573	248	481	1,128,830,151	457,325,999
Berlin - ERDF	2,315	1,712	1,161	2,233,457,595	569,531,687
Brandenburg - ERDF	1,556	1,442	772	716,349,376	573,079,501
Bremen - ERDF	288	285	123	175,468,100	87,734,050
Hamburg - ERDF	67	32	49	169,671,882	84,835,941
Hessen - ERDF	392	384	205	341,560,108	170,780,054
Nordrhein-Westfalen - ERDF	3,047	2,081	1,391	2,411,139,989	1,154,485,565
Saarland - ERDF	235	227	110	133,930,756	62,864,530
Sachsen - ERDF	12,370	7,890	4,712	2,135,346,864	1,708,277,491
Sachsen-Anhalt - ERDF	3,379	1,820	1,421	1,901,310,815	1,463,462,551
Schleswig-Holstein - ERDF	1,033	1,033	717	1,328,156,346	143,622,135
Thüringen - ERDF	4,906	4,900	2,525	1,479,492,698	1,183,594,157
Innovation and Sustainable Growth in Businesses - DK - ERDF	287	273	122	741,311,147	351,224,893
Multi-regional Spain - ERDF	33,012	33,012	12,874	10,077,105,070	6,420,339,500
Andalucía - ERDF	2,710	2,710	1,052	2,632,924,775	2,106,339,820
Aragón - ERDF	486	486	401	173,965,484	86,982,742
Asturias - ERDF	842	842	426	198,483,506	158,786,805
Baleares - ERDF	437	437	203	120,785,277	60,392,638
Canarias - ERDF	1,016	1,016	566	1,102,681,632	937,279,388
Cantabria - ERDF	313	313	187	89,837,032	44,918,516
Castilla y León - ERDF	685	685	311	291,296,346	145,648,173
Castilla-La Mancha - ERDF	1,185	1,185	749	507,493,331	405,994,663
Cataluña - ERDF	1,217	1,217	538	1,342,620,244	671,310,122
Ceuta - ERDF	262	262	221	17,455,522	13,964,418
Valenciana - ERDF	2,136	2,136	1,216	762,482,845	381,241,422
Extremadura - ERDF	6,217	6,217	4,540	745,980,728	596,784,583
Galicia - ERDF	4,015	4,015	2,145	1,024,780,185	819,600,839
La Rioja - ERDF	203	203	110	59,572,324	29,786,162
Madrid - ERDF	394	394	257	127,925,293	63,962,646
Melilla - ERDF	214	214	98	30,724,868	24,579,894
Murcia - ERDF	2,353	2,353	1,019	189,566,832	151,641,486
Navarra - ERDF	470	470	282	50,515,896	25,257,948
País Vasco - ERDF	2,998	2,998	1,893	155,819,517	77,909,758
SME Initiative - ES - ERDF	6	6	1	800,036,600	800,036,600
2014FR (CCI code / programme title unretrievable)	263	258	212	842,317,786	344,921,580
Île-de-France et Seine - ESF/ERDF/YEI	906	871	447	949,457,727	381,305,642
Champagne-Ardenne - ERDF/ESF/YEI	452	422	167	159,723,371	38,106,714
Languedoc-Roussillon - ERDF/ESF/YEI	1,085	1,028	535	776,780,637	328,594,282
Midi-Pyrénées et Garonne - ERDF/ESF/YEI	1,137	1,075	592	1,112,259,898	428,826,359
Martinique - ERDF/ESF/YEI	559	408	450	585,539,476	184,731,477
Bourgogne - ERDF/ESF/YEI	995	978	313	505,492,040	102,690,846
Lorraine et Vosges - ERDF/ESF/YEI	1,262	1,039	688	1,191,845,280	222,058,464
Franche-Comté et Jura - ERDF/ESF	434	428	205	355,827,099	81,073,593
Interregional Pyrénées - ERDF	39	39	31	33,663,892	15,886,397
Alsace - ERDF	208	197	139	88,576,245	27,830,421
Réunion – ERDF	150	149	34	40,416,547	33,961,345
Competitiveness	23,470	22,599	17,750	4,465,568,556	3,496,102,923

Entrepreneurship and Innovation - GR - ERDF/ESF					
Continental Greece - ERDF/ESF	100	88	48	138,344,779	82,110,173
Competitiveness and Cohesion - HR - ERDF/CF	2,666	1,295	1,043	1,591,079,224	1,346,489,505
2014HU (CCI code / programme title unretrievable)	37,557	37,254	20,498	26,366,511,290	19,760,982,256
Border Midland and Western Regional - ERDF	420	387	167	68,206,869	34,103,435
Southern & Eastern Regional Programme - IE - ERDF	352	324	106	109,383,483	54,691,741
2014ITERDF (CCI code / programme title unretrievable)	62,638	55,586	40,892	30,667,988,935	15,409,693,856
2014ITERDFTC (CCI code / programme title unretrievable)	744	503	582	614,516,261	530,180,839
2014ITESF (CCI code / programme title unretrievable)	328,115	287,612	25,292	10,948,072,879	7,374,619,991
Luxembourg - ERDF	22	22	13	45,027,038	18,010,815
Growth and Employment - LV - ERDF/ESF/CF/YEI	1,774	1,768	703	5,081,615,815	3,345,932,062
West Netherlands - ERDF	131	130	116	543,287,536	167,380,714
South Netherlands - ERDF	148	148	126	334,829,117	117,190,191
East Netherlands - ERDF	383	371	346	290,144,447	99,753,730
Dolnośląskie Voivodeship - ERDF/ESF	211	174	72	652,676,535	547,906,533
Łódzkie Voivodeship - ERDF/ESF	234	215	67	863,433,639	575,372,839
Małopolskie Voivodeship - ERDF/ESF	236	234	98	827,133,342	601,484,916
Mazowieckie Voivodeship - ERDF/ESF	2,708	2,683	1,651	2,209,799,669	1,633,225,737
Opolskie Voivodeship - ERDF/ESF	1,277	1,273	624	1,128,511,688	844,147,327
Podkarpackie Voivodeship - ERDF/ESF	2,891	2,869	1,540	2,494,279,613	1,970,460,417
Podlaskie Voivodeship - ERDF/ESF	135	123	51	379,698,479	294,395,476
Pomorskie Voivodeship - ERDF/ESF	1,765	1,754	892	2,486,846,665	1,820,943,405
Śląskie Voivodeship - ERDF/ESF	4,396	4,366	1,823	3,954,494,787	3,010,883,189
Świętokrzyskie Voivodeship - ERDF/ESF	7,601	2,033	1,125	3,503,679,187	2,792,350,113
Warmińsko-Mazurskie Voivodeship - ERDF/ESF	3,831	3,774	1,770	2,008,186,811	1,569,491,993
Wielkopolskie Voivodeship - ERDF/ESF	3,303	3,068	1,742	3,172,975,809	2,089,164,148
Zachodniomorskie Voivodeship - ERDF/ESF	1,523	1,519	814	958,181,030	689,714,508
Smart growth - PL - ERDF	6,327	6,310	4,941	12,265,584,725	7,387,551,235
Digital Poland - ERDF	471	464	185	2,911,769,897	2,404,508,728
Development of Eastern Poland - ERDF	1,156	1,152	923	2,393,046,203	1,884,306,951
Sustainability and Resource Use Efficiency - PT - CF	1,894	1,889	560	2,654,301,107	2,009,860,794
Norte - ERDF/ESF	24,403	7,036	4,656	14,113,882,436	8,151,045,627
Alentejo - ERDF/ESF	3,796	2,930	1,718	1,373,370,572	1,125,145,952
Lisboa - ERDF/ESF	3,005	2,362	1,482	1,670,356,612	779,022,565
Madeira - ERDF/ESF	3,459	304	1,420	914,137,894	306,696,851
Algarve - ERDF/ESF	970	780	0	491,300,334	266,067,145
2014RO (CCI code / programme title unretrievable)	2,261	1,999	1,901	35,172,191,617	27,390,810,045
EU Cohesion Policy - SI - ERDF/ESF/CF/YEI	4,944	4,262	2,726	3,711,232,458	3,089,348,371
2014SK (CCI code / programme	8,607	8,243	4,568	10,939,129,057	8,417,950,495

title unretrievable)					
Quality of Environment - SK - ERDF/CF	568	550	432	1,232,019,255	1,013,180,604
Integrated Regional Programme - SK - ERDF	2,098	2,085	1,366	1,804,099,608	1,725,154,111
2014TC16I5CB005 (CCI code / programme title unretrievable)	78	78	78	21,693,818	18,439,745
2014TC16I5CB006 (CCI code / programme title unretrievable)	65	65	65	15,837,271	13,461,681
2014TC16I5CB007 (CCI code / programme title unretrievable)	68	68	61	26,558,703	22,574,897
2014TC16I5CB009 (CCI code / programme title unretrievable)	40	40	39	38,738,854	32,928,026
Interreg V-B - Mediterranean	130	130	112	284,468,965	240,345,936
Interreg V-B - Adriatic-Ionian	35	35	35	54,934,497	39,196,432
Interreg V-B - Balkan-Mediterranean	43	43	43	41,346,183	28,647,958
Interreg V-A - Belgium-Germany-The Netherlands (Euregio Maas-Rijn)	688	41	251	278,992,521	139,276,741
Interreg V-A - Austria–Germany/Bayern	76	75	75	68,868,391	51,773,643
Interreg V-A - Spain-France-Andorra (POCTEFA)	357	58	310	111,271,164	71,186,036
Interreg V-A - Spain-Portugal (Madeira-Açores-Canarias (MAC))	56	56	56	63,840,437	54,264,372
Interreg V-A - Hungary-Croatia	55	55	50	43,281,545	36,789,313
Interreg V-A - Germany/Bayern-Czech Republic	343	263	214	105,892,529	89,075,500
Interreg V-A - Finland-Estonia-Latvia-Sweden (Central Baltic)	765	127	574	161,276,650	125,920,000
Interreg V-A - Germany (Mecklenburg-Vorpommern-Brandenburg) -Poland	41	41	31	126,444,051	99,221,484
Interreg V-A - Germany-The Netherlands	1,704	168	1,282	453,415,168	227,340,633
Interreg V-A - France-Italy (ALCOTRA)	319	58	271	50,478,281	42,882,615
Interreg V-A - France-Belgium-The Netherlands-United Kingdom (Two seas)	864	82	662	435,348,150	255,758,789
Interreg V-A - France-Germany-Switzerland (Rhin supérieur)	130	130	88	160,874,755	81,970,329
Interreg V-A - France-Switzerland	100	100	73	128,646,563	45,046,436
Interreg V-A - France-Belgium-Germany-Luxembourg (Grande Région)	50	50	44	178,163,842	95,309,427
Interreg V-A - Belgium-The Netherlands	81	81	62	332,520,141	156,425,748
Interreg V-A - United Kingdom-Ireland (Ireland-Northern Ireland-Scotland)	32	28	22	282,495,550	217,319,311
Interreg V-A - United Kingdom-Ireland (Ireland-Wales)	20	20	13	87,560,000	69,520,000
Interreg V-A - Estonia-Latvia	211	29	191	28,904,053	24,351,756
Interreg V-A - Slovenia-Hungary	13	13	13	11,334,111	9,633,994
Interreg V-A - Slovenia-Austria	40	40	35	49,558,853	42,125,025
Interreg V-A - Greece-Cyprus	31	30	24	48,772,461	39,140,712
Interreg Europe	2,089	258	1,642	323,857,919	272,695,002
PEACE (IE-UK)	95	68	58	255,962,214	212,563,486
Interreg V-B - Atlantic Area	45	45	40	117,308,499	87,938,875
Interreg V-B - Central Europe	85	85	81	193,011,515	160,199,558

Interreg V-B - North West Europe	978	95	875	616,046,166	352,528,319
Interreg V-B - South West Europe	480	59	394	94,790,409	71,389,794
Interreg V-B - Indian Ocean Area	150	149	34	40,416,547	33,961,344
2014UK (CCI code / programme title unretrievable)	252	234	68	4,066,247,980	2,668,265,845
England - ERDF	1,247	1,217	519	8,254,354,175	4,176,513,776
Gibraltar - ERDF	73	71	46	29,189,010	11,720,503
Scotland - ERDF	222	216	41	1,435,436,385	633,445,238

Figure 6.: general summary per CCI programme

TYPES OF GEOLOCATIONAL DATA AT THE WHOLE DATASET

The following table contains the categorisation of detected geolocational data according to the NUTS - LAU statistical system.

Please note: the exact validation of each value and categorisation match was out of the scope of this project. It is highly likely that the categorisation contains mismatches as there was a large variance of values even within one dataset (multiple members of different hierarchies were present in a single column). The following summary is rather a general overview of how the distribution of values looks at high level, without dedicating further efforts to validate each values' match in the hierarchy. The mapping of these columns is documented in the source code of the project and is a subject to improvements, additional supervision.

Geolocation level	Operation count	Present	Percentage
NUTS1 code	716,693	0	0
NUTS1 name	716,693	7,190	1
NUTS2 code	716,693	391,719	54.66
NUTS2 name	716,693	413,857	57.75
NUTS3 code	716,693	429,478	59.92
NUTS3 name	716,693	565,630	78.92
LAU1 code	716,693	0	0
LAU1 name	716,693	35,605	4.97
LAU2 code	716,693	397,847	55.51
LAU2 name	716,693	530,529	74.02
City	716,693	10,308	1.44
District	716,693	0	0
Postal code	716,693	118,430	16.52

Figure 7.: distribution of geolocational data on the whole dataset

TYPES OF GEOLOCATIONAL DATA PER NATIONAL DATASET

The distribution of geolocational data per national dataset is the following.

Please note: for easier readability, we only included those hierarchical levels, which had at least one member in the respective national dataset (rows with 0 Present values were excluded).

Country	Geolocation level	Operation count	Present	Percentage
AT	LAU2 name	1,264	1,264	100
BE	City	3,959	203	5.13
BE	Postal code	3,959	3,239	81.81
BG	City	2,634	2,634	100
CZ	NUTS3 code	33,477	33,474	99.99
CZ	NUTS3 name	33,477	33,474	99.99
CZ	Postal code	33,477	33,468	99.97
DE	NUTS1 name	41,670	235	0.56
DE	NUTS3 name	41,670	23,878	57.3
DE	LAU2 code	41,670	3,047	7.31
DE	LAU2 name	41,670	18,781	45.07
DE	City	41,670	1,481	3.55
DE	Postal code	41,670	8,832	21.2
DK	Postal code	287	287	100
ES	NUTS3 name	61,171	61,165	99.99
ES	LAU2 name	61,171	61,001	99.72
ES	Postal code	61,171	27,675	45.24
FR	NUTS2 name	7,490	476	6.36
FR	NUTS3 code	7,490	3,529	47.12
FR	NUTS3 name	7,490	1,293	17.26
FR	LAU2 name	7,490	2,700	36.05
FR	City	7,490	383	5.11
FR	Postal code	7,490	3,254	43.44
GR	Postal code	23,570	23,470	99.58
HR	NUTS2 name	2,666	2,666	100
HR	NUTS3 name	2,666	2,666	100
HU	LAU2 name	37,557	37,309	99.34
IE	NUTS2 name	772	352	45.6
IE	NUTS3 name	772	420	54.4
IE	Postal code	772	352	45.6
IT	NUTS2 code	391,497	391,497	100
IT	NUTS2 name	391,497	391,497	100
IT	NUTS3 code	391,497	391,497	100
IT	NUTS3 name	391,497	391,497	100
IT	LAU2 code	391,497	391,497	100
IT	LAU2 name	391,497	391,497	100
LU	LAU2 name	22	22	100
LV	NUTS3 name	1,774	1,774	100
NL	NUTS2 name	662	148	22.36
NL	LAU2 name	662	661	99.85
NL	Postal code	662	508	76.74
PL	NUTS3 name	38,065	10,559	27.74
PL	LAU1 name	38,065	211	0.55
PL	LAU2 code	38,065	3,303	8.68
PL	LAU2 name	38,065	6,660	17.5
PL	City	38,065	3,303	8.68
PT	NUTS2 name	37,527	15,406	41.05
PT	NUTS3 name	37,527	22,470	59.88
PT	LAU1 name	37,527	35,394	94.32
PT	LAU2 name	37,527	4,569	12.18
PT	Postal code	37,527	3,459	9.22
RO	NUTS1 name	2,261	1,884	83.33
RO	NUTS2 name	2,261	1,884	83.33
RO	NUTS3 name	2,261	1,001	44.27

SI	NUTS1 name	4,944	4,941	99.94
SI	NUTS3 name	4,944	4,944	100
SI	LAU2 name	4,944	4,944	100
SI	Postal code	4,944	4,944	100
SK	NUTS3 name	11,273	8,532	75.69
SK	LAU2 name	11,273	568	5.04
TC	NUTS1 name	10,357	130	1.26
TC	NUTS2 name	10,357	1,176	11.35
TC	NUTS3 code	10,357	978	9.44
TC	NUTS3 name	10,357	1,957	18.9
TC	LAU2 name	10,357	553	5.34
TC	City	10,357	2,304	22.25
TC	Postal code	10,357	7,473	72.15
UK	NUTS2 code	1,794	222	12.37
UK	NUTS2 name	1,794	252	14.05
UK	Postal code	1,794	1,469	81.88

Figure 8.: distribution of geolocational data per national dataset

PRESENCE OR ABSENCE OF ADDITIONAL, NON REQUIRED COLUMNS AT THE WHOLE DATASET

The scope of the project also included the detection of non required categorisation systems. These proved to be rather sparse and messy: at times it involved guesswork the assigning which column in the source data translates to which concept. The mapping of these columns is documented in the source code of the project and is a subject to improvements, additional supervision.

Column name	Transaction count	Present	Percentage (%)
Beneficiary ID	716,693	463,095	64.62
Operation ID	716,693	528,621	73.76
Priority axis	716,693	519,520	72.49
Form of finance	716,693	490,618	68.46
Territorial dimension	716,693	104,537	14.59
Territorial delivery mechanism	716,693	86,068	12.01
ESF secondary theme	716,693	63,668	8.88
Economic dimension	716,693	7,139	1

Figure 9.: presence of additional columns on the whole dataset

PRESENCE OR ABSENCE OF ADDITIONAL, NON REQUIRED COLUMNS PER NATIONAL DATASET

The distribution of the additional fields across national dataset is listed in the next table.

Please note: for easier readability, we only included those columns, which had at least one value mapped to in the respective national dataset (rows with 0 Present values were excluded).

Country	Column name	Transaction count	Present	Percentage (%)
BE	Operation ID	3,959	923	23.31
CZ	Beneficiary ID	33,477	33,451	99.92

CZ	Operation ID	33,477	33,477	100
CZ	Priority axis	33,477	33,477	100
DE	Operation ID	41,670	19,476	46.74
DE	Form of finance	41,670	640	1.54
DE	Territorial dimension	41,670	639	1.53
DE	Territorial delivery mechanism	41,670	640	1.54
DE	ESF secondary theme	41,670	874	2.1
DE	Economic dimension	41,670	67	0.16
DK	Priority axis	287	287	100
DK	Territorial dimension	287	285	99.3
DK	Territorial delivery mechanism	287	285	99.3
DK	ESF secondary theme	287	19	6.62
ES	Form of finance	61,171	61,159	99.98
ES	Territorial dimension	61,171	61,150	99.97
ES	Territorial delivery mechanism	61,171	60,740	99.3
GR	Priority axis	23,570	23,570	100
GR	Territorial dimension	23,570	100	0.42
HU	ESF secondary theme	37,557	37,557	100
IE	Operation ID	772	352	45.6
IT	Beneficiary ID	391,497	391,497	100
IT	Operation ID	391,497	391,497	100
IT	Priority axis	391,497	391,497	100
IT	Form of finance	391,497	391,497	100
LU	Operation ID	22	22	100
LU	Priority axis	22	22	100
LV	Beneficiary ID	1,774	1,774	100
LV	Operation ID	1,774	1,774	100
LV	Priority axis	1,774	1,774	100
NL	Operation ID	662	148	22.36
NL	Priority axis	662	662	100
PL	Operation ID	38,065	30,251	79.47
PL	Priority axis	38,065	26,948	70.79
PL	Form of finance	38,065	26,343	69.21
PL	Territorial dimension	38,065	26,343	69.21
PL	ESF secondary theme	38,065	25,187	66.17
PT	Beneficiary ID	37,527	27,862	74.25
PT	Operation ID	37,527	35,633	94.95
PT	Priority axis	37,527	18,712	49.86
PT	Form of finance	37,527	10,757	28.66
PT	Territorial dimension	37,527	15,798	42.1
PT	Territorial delivery mechanism	37,527	24,403	65.03
PT	Economic dimension	37,527	6,801	18.12
RO	Operation ID	2,261	2,210	97.74
RO	Priority axis	2,261	1,990	88.01
SI	Operation ID	4,944	4,944	100
SI	Priority axis	4,944	4,944	100
SK	Beneficiary ID	11,273	8,511	75.5
SK	Operation ID	11,273	568	5.04
SK	Priority axis	11,273	8,533	75.69
TC	Operation ID	10,357	7,064	68.21
TC	Priority axis	10,357	5,383	51.97
TC	ESF secondary theme	10,357	31	0.3
TC	Economic dimension	10,357	49	0.47
UK	Operation ID	1,794	282	15.72
UK	Priority axis	1,794	1,721	95.93
UK	Form of finance	1,794	222	12.37
UK	Territorial dimension	1,794	222	12.37
UK	Economic dimension	1,794	222	12.37

Figure 10.: presence of additional columns per national dataset

DATA PUBLISHING GUIDELINES

Based on the formats, structure and data types encountered during the pilot project, I recommend following a guideline, which ensures that the data published across multiple member states and regions is more accessible, standardized and contains less bias when processed automatically.

A. Preferred file format

Data should be published in .csv (comma separated value) file formats in text form. The .csv file should follow the RFC 4180 standard (https://tools.ietf.org/html/rfc4180). Text fields should be double quoted, the default delimiter should be comma.

B. Default encoding

Data published in .csv file formats should always be encoded in UTF-8 character encoding. If this is not possible, the used encoding should always be stated. This can be included as text in a different file bundled together with the .csv file, or on the first row of the .csv file.

C. Include header

The header row should be the first row of the .csv file - if the data is not in UTF-8 encoding, and the encoding information is the first row, then the header should be the second row. All columns / fields should have a unique name, and should be descriptive, preferably in English. A column name should not be longer than 355 characters.

D. Files with multiple sheets

The .csv file format does not allow for multiple tabs or datasheets within one file. Therefore if there are multiple sheets, they should be exported in separate .csv files or if they have the same data structure, merged together into one .csv file.

E. Merged cells and number of columns

The .csv file format also does not allow merged cells - merging cells within one row, or merging several rows within one column should be avoided at all times. In other words the file should contain exactly the same number of columns at each row as the header.

F. Subtotal rows, total rows

No aggregations should be included in the dataset: each row should only contain a single transaction or operation, and no aggregation of other transactions or operations.

G. Data types

We recommend following these formatting styles:

- Dates should be formatted following the YYYY-MM-DD formatting style.
- Time should be formatted following the YYYY-MM-DD HH24:MI:SS.MS formatting style.
- Integers should always be expressed as integers. No thousand separators should be included (space, comma), only 0-9 integer values.
- Numeric values should only contain 0-9 integer values separated by a dot (.) No thousand separators (space, comma) should be included.
- In case a numeric value is a currency value, the currency symbol (for example € or $) should not be included with the value, but included in a separate field.
- Currency codes should follow the ISO 4217 standard (three character currency codes).
- Text should always be double quoted.

H. CCI codes

CCI codes should be included on each row.

I. Publishing on a persistent URL

The most recent data should always be accessible on a persistent web address: the url should not change according to different versions, updates of the file. Therefore there should always be a master copy available, containing all transactions to date, accessible on a predefined URL.

APPENDIX A

REPORTS ON ACCESSING SOURCE DATA

INTRODUCTION

Appendix A contains the rather technical tables, which summarise details about accessing the source data and the state and shape of the source data before loading it into a database. We felt it necessary to include in the final report as they contain vital information about reasons why certain datapoints might be missing from the final master database.

PROGRAMME CODES WITH NO ENDPOINTS

The following programme codes were missing endpoints, therefore a file could not be accessed:

Country code	Programme code
CY	2014CY16M1OP001
FI	2014FI05M2OP001
FI	2014FI16M2OP001
FI	2016FI16RFSM001
GR	2014GR05M2OP001
GR	2014GR16M2OP002
GR	2014GR16M2OP003
GR	2014GR16M2OP004
GR	2014GR16M2OP005
GR	2014GR16M2OP006
GR	2014GR16M2OP008
GR	2014GR16M2OP009
GR	2014GR16M2OP010
GR	2014GR16M2OP011
GR	2014GR16M2OP012
GR	2014GR16M2OP013
GR	2014GR16M2OP014
GR	2014GR16M3TA001
MT	2014MT16M1OP001
MT	2014MT16RFSM001
NL	2014NL16RFOP001
PL	2014PL16M2OP002
PL	2014PL16M2OP003
PL	2014PL16M2OP004
PT	2014PT16M2OP002
PT	2014PT16M2OP004
PT	2014PT16RFTA001
SE	2014SE16M2OP001
SE	2014SE16RFOP001
SE	2014SE16RFOP002
SE	2014SE16RFOP003
SE	2014SE16RFOP004
SE	2014SE16RFOP005
SE	2014SE16RFOP006
SE	2014SE16RFOP007
SE	2014SE16RFOP008
SE	2014SE16RFOP009

SK	2014SK05M0OP001
SK	2014SK16M1OP001
SK	2014SK16RFOP001
SK	2014SK16RFTA001
TC	2014TC16I5CB001
TC	2014TC16I5CB002
TC	2014TC16I5CB003
TC	2014TC16I5CB004
TC	2014TC16I5CB008
TC	2014TC16I5CB010
TC	2014TC16M5TN001
TC	2014TC16M6TN001
TC	2014TC16RFCB002
TC	2014TC16RFCB003
TC	2014TC16RFCB005
TC	2014TC16RFCB010
TC	2014TC16RFCB011
TC	2014TC16RFCB012
TC	2014TC16RFCB013
TC	2014TC16RFCB016
TC	2014TC16RFCB018
TC	2014TC16RFCB020
TC	2014TC16RFCB021
TC	2014TC16RFCB022
TC	2014TC16RFCB024
TC	2014TC16RFCB026
TC	2014TC16RFCB027
TC	2014TC16RFCB028
TC	2014TC16RFCB030
TC	2014TC16RFCB032
TC	2014TC16RFCB033
TC	2014TC16RFCB035
TC	2014TC16RFCB036
TC	2014TC16RFCB037
TC	2014TC16RFCB040
TC	2014TC16RFCB042
TC	2014TC16RFCB043
TC	2014TC16RFCB044
TC	2014TC16RFCB049
TC	2014TC16RFCB051
TC	2014TC16RFCB052
TC	2014TC16RFCB056
TC	2014TC16RFIR002
TC	2014TC16RFIR003
TC	2014TC16RFIR004
TC	2014TC16RFTN001
TC	2014TC16RFTN004
TC	2014TC16RFTN005
TC	2014TC16RFTN008
TC	2014TC16RFTN010
UK	2014UK16RFOP003

Figure 1.: programme codes with missing endpoints

ENDPOINTS WITH MULTIPLE PROGRAMME CODES

The following programme codes were associated with the same endpoint, therefore a 1:1 relationship could not be established between the two.

Country code	Unique programme code	Programme code array
CZ	2014CZ	{2014CZ16RFOP001,2014CZ16RFOP002,2014CZ16M1OP001,2014CZ16M1OP002,2014CZ16M2OP001,2014CZ16CFTA001,2014CZ05M2OP001}
FR	2014FR	{2014FR16RFOP001,2014FR16M0OP013}
FR	2014FR	{2014FR05M2OP001,2014FR16M0OP009}
FR	2014FR	{2014FR16M0OP008,2014FR16M0OP012}
FR	2014FR	{2014FR16M2OP009,2014FR16M2OP006,2014FR16M0OP001}
FR	2014FR	{2014FR16RFOP002,2014FR16M2OP003,2014FR16M2TA001,2014FR16M0OP002,2014FR16M0OP003,2014FR16M0OP005,2014FR16M2OP001,2014FR16M2OP004,2014FR16M2OP008,2014FR16M2OP010,2014FR16M2OP011,2014FR16M2OP012,2014FR16RFOP003,2014FR16RFOP005}
HU	2014HU	{2014HU16M2OP001,2014HU16M0OP001,2014HU16M1OP001,2014HU16M1OP003,2014HU05M3OP001,2014HU16M2OP002,2014HU05M2OP001}
PL	2014PL	{2014PL16CFTA001,2014PL16M1OP001}
RO	2014RO	{2014RO16RFOP001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}
UK	2014UK	{2014UK16RFOP005,2014UK16RFOP006}

Figure 2.: programme codes with shared endpoints

UNSUCCESSFUL FILE DOWNLOADS

An error was repeatedly encountered while trying to access the following endpoints, therefore the files could not be retrieved:

Country code	Programme code	Endpoint
GR	2014GR16M1OP001	https://www.espa.gr/el/Documents/ListOfOperations_20200128.zip

Figure 3.: unsuccessful file downloads

UNSUCCESSFUL UNCOMPRESSIONS

All files were successfully uncompressed, we encountered no errors.

UNSUCCESSFUL FILE CONVERSIONS AND CORRUPTED FILES

The following downloaded files were corrupted to the extent that their contents could not be accessed:

Country code	Programme code	Endpoint

TC	2014TC16RFCB015	http://www.skhu.eu/_projectxls
TC	2014TC16RFCB017	https://www.sn-cz2020.eu/media/de_cs/einzelne_dokumente/20-03-19_Liste_der_Vorhaben.xlsx
TC	2014TC16RFCB025	http://pl.cz-pl.eu/obsah/soubory/734/ogolna-lista-zatwierdzonych-projektow-do-24.06.2019.xlsx
TC	2014TC16RFCB029	http://pl.cz-pl.eu/ogoszenia-o-naborach-wnioskow-i-zatwierdzone-projekty

Figure 4.: corrupted files

UNSUCCESSFUL CHARACTER ENCODINGS

Character encoding threw some errors where the source was in .csv format and the file contained non UTF-8 characters. We decided to discard these characters, thus saving the rest of the file contents. The extent of truncation was not significant throughout the processing of the national datasets.

FILES WITH NO HEADER OR DETECTABLE TABULAR STRUCTURE

For the following programme codes and files it was either not possible to detect a valid header or the data could not be converted to a tabular structure. The file sheet index following the underscore appended to the programme codes indicate the worksheet's index within the original source file workbook, or if there's multiple files belonging to the same programme code, their index according to alphabetical order. Most cases included here are simply not tabular data: the worksheet within the workbook contains some sort of description or free text addition to the data.

Country code	Unique programme code	Programme code array	Endpoint	File sheet index
CZ	2014CZ	{2014CZ16RFOP001,2014CZ16RFOP002,2014CZ16M1OP001,2014CZ16M1OP002,2014CZ16M2OP001,2014CZ16CFTA001,2014CZ05M2OP001}	https://www.dotaceeu.cz/getmedia/c654292a-a424-428e-95f1-4c28baccd7a9/2019_03_01-M023a-Seznam-operaci-_-List-of-operations_1.xls.aspx?ext=.xls	2
DE	2014DE16RFOP005	{2014DE16RFOP005}	https://www.efre-bremen.de/sixcms/media.php/13/2019-12-31_Liste%20der%20Vorhaben.xlsx	2
EE	2014EE16M3OP001	{2014EE16M3OP001}	https://www.struktuurifondid.ee/eng/toetatud-projektid/toetatud_projektid.csv	1
GR	2014GR16M2OP007	{2014GR16M2OP007}	http://stereaellada.gr/wp-content/uploads/2018/02/%CE%9A%CE%91%CE%A4%CE%91%CE%9B%CE%9F%CE%93%CE%9F%	2

			CE%A3-%CE%A0%CE%A1%CE%91%CE%9E%CE%95%CE%A9%CE%9D-%CE%99%CE%91%CE%9D%CE%9F%CE%A5%CE%91%CE%A1%CE%99%CE%9F%CE%A3-2018-1.xls	
GR	2014GR16M2OP007	{2014GR16M2OP007}	http://stereaellada.gr/wp-content/uploads/2018/10/%CE%92-%CE%95%CE%9E%CE%91%CE%9C%CE%97%CE%9D%CE%9F-2018-%CE%9F%CE%B9%CE%BA%CE%BF%CE%BD%CE%BF%CE%BC%CE%B9%CE%BA%CE%AC-%CF%83%CF%84%CE%BF%CE%B9%CF%87%CE%B5%CE%AF%CE%B1-%CE%AD%CF%81%CE%B3%CF%89%CE%BD-1.xlsx	2
HU	2014HU	{2014HU16M2OP001,2014HU16M0OP001,2014HU16M1OP001,2014HU16M1OP003,2014HU05M3OP001,2014HU16M2OP002,2014HU05M2OP001}		2
PL	2014PL16M2OP007	{2014PL16M2OP007}	https://www.funduszedlamazowsza.eu/wp-content/uploads/2019/03/lista-projektow-konkursowych.xlsx	2
PT	2014PT16M2OP007	{2014PT16M2OP007}	https://algarve2020.eu/info/sites/algarve2020.eu/files/candidaturas/20200414_quadro_ii_operacoes_aprovadas_31-03-2020.xlsx	2
PT	2014PT16M2OP007	{2014PT16M2OP007}	https://algarve2020.eu/info/sites/algarve2020.eu/files/candidaturas/20200414_quadro_ii_operacoes_aprovadas_31-03-2020.xlsx	3
RO	2014RO	{2014RO16RFOP001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}	http://www.fonduri-ue.ro/images/files/implementare-absorbtie/2017/Lista_Proiecte_contractate_-_01.2018.zip	4
RO	2014RO	{2014RO16RFO	http://www.fonduri-	11

		P001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}	ue.ro/images/files/implementare-absorbtie/2017/Lista_Proiecte_contractate_-_01.2018.zip	
RO	2014RO	{2014RO16RFOP001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}	http://www.fonduri-ue.ro/images/files/implementare-absorbtie/2017/Lista_Proiecte_contractate_-_01.2018.zip	14
RO	2014RO	{2014RO16RFOP001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}	http://www.fonduri-ue.ro/images/files/implementare-absorbtie/2017/Lista_Proiecte_contractate_-_01.2018.zip	15
RO	2014RO	{2014RO16RFOP001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}	http://www.fonduri-ue.ro/images/files/implementare-absorbtie/2017/Lista_Proiecte_contractate_-_01.2018.zip	16
RO	2014RO	{2014RO16RFOP001,2014RO16M1OP001,2014RO16RFOP002,2014RO16RFTA001,2014RO16RFSM001}	http://www.fonduri-ue.ro/images/files/implementare-absorbtie/2017/Lista_Proiecte_contractate_-_01.2018.zip	17
TC	2014TC16RFCB006	{2014TC16RFCB006}	https://www.poctefa.eu/wp-content/uploads/2018/05/Listado-proyectos-programados_FR_180518.xlsx	2
TC	2014TC16RFCB039	{2014TC16RFCB039}	https://www.interreg-rhin-sup.eu/wp-content/uploads/projets-acceptes-angenommenen-projekten-12122019-1.xlsx	7
TC	2014TC16RFCB047	{2014TC16RFCB047}	http://seupb.eu/sites/default/files/styles/INTERREGVA/Jun_Beneficiaries_Spreadsheet_INTERREG_Webcopy_English.xlsx	2
TC	2014TC16RFPC001	{2014TC16RFPC001}	http://seupb.eu/sites/default/files/styles/PEACEIV/Apr_Beneficiaries_Spreadsheet_PEACE_Webcopy_English_0.XLSX	2
UK	2014UK	{2014UK16RFOP005,2014UK16RFOP006}	https://gov.wales/sites/default/files/publications/2020-02/eu-structural-funds-programme-2014-	2

			to-2020-approved-projects.ods	
UK	2014UK16RFOP004	{2014UK16RFOP004}	https://www.gov.scot/binaries/content/documents/govscot/publications/transparency-data/2018/06/esif-operations-funding/documents/esf-and-erdf-operations-funding-approved-to-july-2019/esf-and-erdf-operations-funding-approved-to-july-2019/govscot%3Adocument/ESF%2Band%2BERDF%2Boperations%2Bfunding%2Bapproved%2Bto%2BJanuary%2B2020.xlsx?forceDownload=true	2

Figure 5.: files with no detectable header or tabular structure

APPENDIX B

METHODOLOGICAL REPORT

INTRODUCTION

This document is a part of deliverables defined by service contract **No. 2020CE16BAT015.** In the following we shall describe the methodologies implemented to meet the requirements of the named contract.

DEVELOPMENT ENVIRONMENT

The development environment of the project consists of the following tools:

- **UNIX shell:** Bash 3.2.57(1)-release
- **Workflow management:** GNU make
- **Database:** PostgreSQL 12.2
- **Accessing and transforming source files:** Python 3.7.5

DEFINITION AND COLLECTION OF ONLINE DATA SOURCES

https://github.com/balkey/eu1420/blob/master/makefile#L131

The very first step was to define the exact website URL's containing the downloadable files for each regional and national cohesion policy programme, since the list of known websites provided by the Commission - as part of the Service contract - contains mostly list pages, where typically a collection of files are available. In some cases these are historical versions of the same files, in other cases these list pages just contain links to further pages containing the files themselves. Further on, we will refer to the websites URL's containing the downloadable files as "endpoints".

In a few cases the **Download url plus readable format** column present in the file provided by the Commission was outdated (the links were broken or moved to a different address within the host), therefore where it was possible, we tried to relocate the the files in their new location.

In addition to collecting the endpoints themselves, the following columns were appended to each endpoint:

- **Access**: a column defining if the endpoint was accessible or not. Used for automated programatic access through scripts.
- **Anchor text:** if the link pointing to the file contained human readable text, it was saved in this column.
- **Last update:** if an anchor text was provided and a recognisable date format could be extracted from the text, it is saved here in YYYY-MM-DD format.
- **File format:** this column specifies the file extension the original source data is stored in, which are used for automated programatic access through scripts. The values can be:
  - csv
  - xlsx (including xls)
  - ods

    - xml
- **compressed:** the boolean value specifies if the source file is compressed or not. Used for automated programatic access through scripts.

The contents of this table were shared with the Commission and went through multiple iterations by assigning a master copy within the CIRCABC portal, before the Commission approved the contents which included their inputs and change requests. The approved copy of the master is accessible at the following URL.

COLLECTION OF OFFLINE SOURCES

https://github.com/balkey/eu1420/blob/master/makefile#L140-L141

Some of the files - e.g. the Hungarian and part of the Slovakian datasets - were not accessible online, however the Commission was able to get copies of the files. These were marked as **offline** in the data source table containing the endpoints and were included manually instead of accessing them programatically through public URL's.

AGGREGATION OF PROGRAMME CODES AND ENDPOINTS

https://github.com/balkey/eu1420/blob/master/database_setup.mk#L16

In some cases, multiple programme codes were assigned to the same endpoint, therefore at this point a 1:1 relation between file and programme code could not be established, in other words it could not be recreated which programme codes belonged to the transactions stored in a specific file, as there were multiple matches. In order to be able to recreate this relationship between a transaction row and a programme code (should there be any kind of reference of the programme code in the processed data), in such cases the programme codes were truncated to the first 6 characters, and the belonging original programme codes (those that shared the same endpoint file) were associated with this truncated version.

This relationship can be expressed with the following example:

country	reference	reference_array	endpoint
PL	2014PL	{2014PL16CFTA001,2014PL16M1OP001}	https://www.pois.gov.pl/media/82768/lista_projektow_POIiS_stan_20191205.xlsx
PL	2014PL	{2014PL16M2OP003,2014PL16M2OP002,2014PL16M2OP004}	
PL	2014PL16M2OP001	{2014PL16M2OP001}	http://rpo.dolnyslask.pl/wp-content/uploads/2020/01/lista-prj-pozak.xlsx

Figure 1.: relationship between endpoints belonging to multiple programme codes

CREATING A FOLDER STRUCTURE ACCORDING TO PROGRAMME CODES

https://github.com/balkey/eu1420/blob/master/makefile#L119

The above specified relations also specify the folder structure we prepared in the local filesystem to download the files to. An example of how the autogenerated folder structure is constructed:

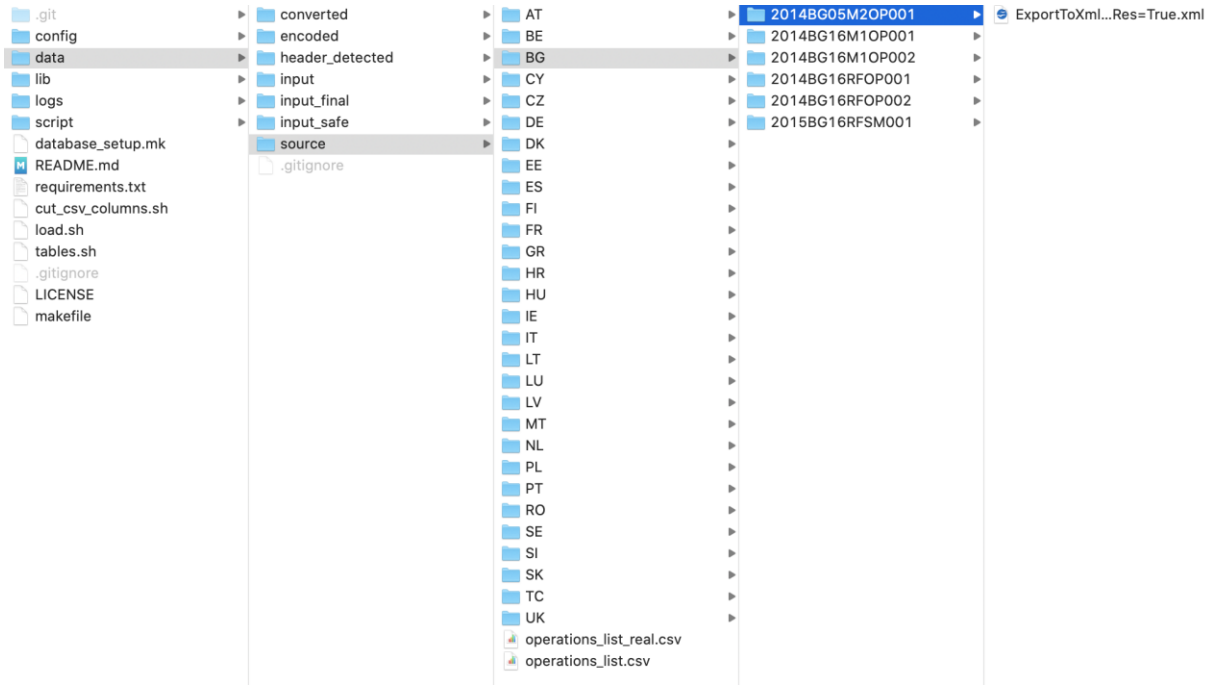

Figure 2.: folder structure autogenerated from programme codes

This structure is maintained in all transformation phases:

1. source
2. converted
3. encoded
4. header detected
5. input (_**safe** for manual edits, _**final** for loading to database)

Under these phases programmes are always separated from each other. In other words, the files are saved and stored after each step of transformation.

The folder structure is dynamically generated with each run, the basis being the contents and structure of the master data table.

DOWNLOADING OF THE FILES

https://github.com/balkey/eu1420/blob/master/makefile#L120

The formerly specified logs are then accessed in a HTTP GET request, and if successful - the request gets a 200 response code - are saved locally with the specified extension in the source **master** table. The reason for this is that the extensions used in the URL's themselves were not always correct, and automatic extension detection based on mime types did not always bring unbiased matches.

If the request was not successful, the error was logged. An example of a logged event of unsuccessful download is the following:

```
2020-06-14 15:40:22,953 - Error occurred with downloading file
from https://www.espa.gr/el/Documents
ListOfOperations_20200128.zip
to folderdata/source//GR/2014GR16M1OP001
with fileformat: csv.
Message: ('Connection aborted.', OSError(0, 'Error'))
```

The files are saved in the corresponding folder, the original filename represented in the endpoint - with the exception of the extension - is maintained locally as well.

UNCOMPRESSION OF FILES

https://github.com/balkey/eu1420/blob/master/makefile#L111

If a file is compressed, the programatic, automated uncompressing involves unnesting the structure within the compressed folder, meaning any nested folder structures are bypassed, and only the files are kept. These files then are moved to their root folder (named after either a programme code or a truncated, six character programme code at this point), and the discarded folders are cleaned up.

FILE CONVERSIONS

https://github.com/balkey/eu1420/blob/master/makefile#L105

In order to be able to load the files into a relational database, first we need to ensure that they are represented in a standard, valid and unified tabular format. The most common structure used for this is a comma separated file, so we will use this standard (RFC4180).

The following steps are included during file conversion:

- If a file is already in .csv format, we move it to the next stage by copying it to the corresponding folder under the **converted** folder branch, without any modification. This is a naive approach, trusting that the .csv file is valid. A possible improvement could be to validate if the rules of the applied .csv standard are respected.
- If a file is an Excel workbook (extension is .xls or .xlsx), we use the **xlrd** open source Python library to access the contents. We save each worksheet in the workbook as a separate file, and to maintain their original order, we include their index in the filename. The strings used to identify the worksheet within the workbook (the tab's name) are truncated.
- If the file is an OpenDocument Spreadsheet (extension is .ods), we use the **ezodf** open source Python library to access the contents. The logic is the same as in the case with the .xls files: we save each worksheet separately, and indicate original index order in the filename. The strings used to identify the worksheet within the workbook (the tab's name) are truncated.
- Both the .xls and the .ods parser checks for merged cells. If a merged cell is detected, we take the value from the previous rows corresponding column (from the column index where we are in the present row). This is a first iteration and a

naive approach, therefore it should be improved further, by logging this event in the file itself in a dedicated column. So for each row, we should have a column describing if there was unmerging within the scope of the actual row, and another column storing the number of unmerged cells. This way later on we can at least recall those rows which heavily contained merged cells, and filter them from the end results. A possible use-case could be *Subtotal* rows, which often come with merged cells.

- Both parsers remove line-breaks and newline delimiters from each cell's value in order to ensure a valid .csv structure safely. This transformation somewhat changes the end results (for example a long project description containing multiple paragraphs will be truncated into a single paragraph), but this we treated as a representational and not content specific issue.
- If the source file is stored in an .xml format, we simply flatten it to a comma separated format and move it to the next transformation stage. The so far detected .xml files only contained data only in the following structure:

```
<Table>
  <Row>
    <Cells>
      <Cell>
        <Value>
          Example string
        </Value>
      </Cell>
    </Cells>
  </Row>
</Table>
```

Therefore a more complex, recursive object parsing was not necessary and was not yet implemented.

As with every step, we log the errors in this step. An example log-line looks like:

```
2020-06-17 16:31:43,192 - Error occurred with
./data/source/TC/2014TC16RFCB017/
20-03-19_Liste_der_Vorhaben.xlsx.xlsx
Message: Unsupported format, or corrupt file:
Expected BOF record; found b'\n\n<!DOCT'
```

ENSURE UTF-8 ENCODING

https://github.com/balkey/eu1420/blob/master/makefile#L100

In this step we try to encode all the contents into UTF-8 character encoding. If a character can't be converted to UTF-8, we discard it from the output of this step. This problem only appears with files where the source is stored in .csv format.

Many iterations were spent on automated character encoding detection, but with unsatisfactory results so far. Generally, if the encoding of the source file is not provided, it is more or less guesswork to try to find out which encoding is used. Since so far most of the files seem to retain their original contents, we accepted ignoring encoding errors and discarding of the unconvertible characters as a good enough solution. Some of the Greek files are exceptions, there most of the contents had to be discarded. An improvement could be to provide the most probable encodings with each file based on

their original language which can be detected based on the country codes in the programme codes.

DELIMITER DETECTION

https://github.com/balkey/eu1420/blob/master/lib/encode.py#L45

In every use-case where the files are parsed - from .xls, .ods or .xml formats - the delimiter is always a comma. However, some endpoints where data is stored in .csv format, there are other characters used for delimiting the columns, the semi-colon being one example in the Italian dataset.

For performance issues, we take the first 2 megabytes of the file available, and try to guess the most probable candidate for a delimiter. Then we standardise the file and ensure that the delimiter is always a comma.

If for some reason, a delimiter can not be automatically detected, we log the event. An example log-line looks like:

```
2020-06-17 16:34:00,501 - Error occurred with delimiter
detecting file
data/converted/TC/2014TC16RFPC001
Apr_Beneficiaries_Spreadsheet_PEACE_Webcopy_English_
0.XLSX.xlsx_2_Sheet1.csv from folder ""
Approved by SC
Draft LoO Issued
Final LoO Issued
LoO Accepted .
Message: Could not determine delimiter
```

A possible reason for not being able to detect a delimiter is that the contents of the given spreadsheet are not actually tabulated. We have seen use-cases where charts or free text is included in the spreadsheets. These would have to be manually examined based on the provided logs.

REMOVAL OF EMPTY FILES

https://github.com/balkey/eu1420/blob/master/makefile#L106

Some of the workbooks contain empty worksheets: they are created in the workbook, likely have a name, but contain no data. At this step, we clean these files up and delete them if such an empty file was created with no data in it.

AUTOMATED HEADER DETECTION AND FILE META ATTRIBUTES LOGGING

https://github.com/balkey/eu1420/blob/master/makefile#L95

In this step, we try to detect the header line in the prepared files, since the header is not necessary contained on the first line - before the actual header many files have incomplete or empty rows, or some rows contain free text as description of the contents.

For the header detection we applied the following logic:

- Starting from row one, we count the width (number of detected columns) of the row, excluding columns with empty string or NULL values. This integer will be our candidate for the width of the complete tabular data. We save this row as a possible candidate for header in a separate .json file.
- Then we iterate over all the rows of the file. If we encounter a row where the possible width - excluding empty rows - is greater than the formerly selected candidates' width, we save this candidate to the .json output, and set the possible width of the tabular data to this integer.
- This iteration is continued until the last row of the file.
- If we encounter a column with a NULL value or empty string in the possible candidates, we assign a **missing_column_name** value to it, and append its index number among the rows columns which had NULL or empty value so far, so the second appearance of an empty column name for example becomes **missing_column_name_2**. This is to ensure that all missing columns with missing values have unique names in the possible header candidate list.
- It is important that with this method, the original column order is maintained!
- In the preserved json object we also log the index (row number) of the header candidate within the file. We also store the current detected width belonging to the candidate.
- A meta log like the following example is saved for each file:


```
[
    {
        "content": [
            "Liste der Vorhaben / list of operations",
            "missing_column_name_1",
            "missing_column_name_2",
            "missing_column_name_3",
            "missing_column_name_4",
            "missing_column_name_5",
            "missing_column_name_6",
            "missing_column_name_7",
            "missing_column_name_8",
            "missing_column_name_9",
            "missing_column_name_10",
            "missing_column_name_11",
            "missing_column_name_12",
            "missing_column_name_13"
        ],
        "content_width": 14,
        "row_number": 0
    },
    {
        "content": [
            "Name des Begünstigten / beneficiary name",
            "Bezeichnung des Vorhabens / operation name",
            "Zusammenfassung des Vorhabens / operation summary",
            "Datum des Beginns des Vorhabens / operation start date",
            "Datum des Endes des Vorhabens / operation end date",
            "Gesamtbetrag der förderfähigen Ausgaben des Vorhabens /
total eligible expenditure allocated to the operation",
            "Unions-Kofinanzierungssatz pro Prioritätsachse / Union
co-financing rate, as per priority axis",
            "Durchführungsort / location",
            "Land / country",
            "Interventionskategorie / name of category of
```

```
      intervention",
                  "missing_column_name_1",
                  "missing_column_name_2",
                  "missing_column_name_3",
                  "missing_column_name_4"
              ],
              "content_width": 14,
              "row_number": 5
          },
          {
              "content": [
                  "A & L StoneROB OHG",
                  "Erschließung des Zielmarktes USA",
                  "Export von Produkten in die USA und Anpassung der
      Webseite und Prospekte auf den amerikanischen Markt. ",
                  "43790.0",
                  "44155.0",
                  "40000.0",
                  "25%*",
                  "Kirchberg , Wald",
                  "Deutschland",
                  "066 - Fortgeschrittene Unterstützungsdienste für KMU und
      KMU-Zusammenschlüsse (einschließlich Dienstleistungen für Management,
      Marketing und Design)",
                  "01 - Nicht rückzahlbare Finanzhilfe",
                  "03 - Ländliche Gebiete (dünn besiedelt)",
                  "07 - Nicht zutreffend",
                  "03 - Stärkung der Wettbewerbsfähigkeit von kleinen und
      mittleren Unternehmen"
              ],
              "content_width": 14,
              "row_number": 7
          }
]
```


As it is observable, column names could not always be captured - mainly due to merged cells across multiple rows - but this solution produced a good enough result to proceed to the next step.

AUTOMATED VALIDATION OF SELECTION OF HEADER FROM POSSIBLE CANDIDATES WITH USER INPUT SUPPORT

https://github.com/balkey/eu1420/blob/master/makefile#L85

From the generated .json logs in the previous step, we felt necessary to include a manual supervision of the prepared files. The main argument for this is that a manual / human intelligence driven mapping (translation) of column names is necessary regardless, so there would be a manual step perforce.

To support manual supervision, a script goes through all the generated candidate lists, and through the interactive shells prompts user input to choose which candidate looks the most likely for the final header. The user can choose only one object, which resembles most the headers - even if some destroyed or unrecognisable column names are present. For the user's decision, the row number should be a very good indicator - if it is larger than 10 in our heuristic experience it is likely that we missed the actual header.

Manual supervision is also useful for detecting those prepared files which do actually not contain tabular data. The user also has the chance to assign no header - for these use-cases the .json file will contain an empty array (**[ ]**).

COLUMN NAME TRANSFORMATIONS

https://github.com/balkey/eu1420/blob/master/makefile#L76

Once a possible candidate is selected, we transform the column names so they became valid according to SQL standards. The following transformations are carried out:

- Spaces, line-breaks and tabs are replaced with underscore characters (_).
- Accented characters are replaced with non accented characters.
- Only alphanumerical characters are kept, with the exception of underscore characters. Other characters are escaped.
- Every character is lower case.
- If the column name string after transformations is longer than 63 characters - the SQL column name character limit - we truncate the rest of the string.
- If there are multiple column name values with the same value, we prepend an underscore and an index of the encountered duplication within the array, ensuring that column names are unique - also an SQL requirement.
- The transformed header is also appended to the .json log for later reuse, as a separate object from the original detected header.

ENSURING HEADER IS THE FIRST ROW

https://github.com/balkey/eu1420/blob/master/makefile#L77

Once the final candidate for the header is selected and the column values are transformed to be SQL compatible, we reach back to the **row_number** attribute exposed earlier, and remove all rows where the index is smaller or equal than the value of the attribute. As a next step, we prepend the final header row to the remaining data, thus ensuring that the validated header is the first row.

REMOVAL OF EMPTY COLUMNS

https://github.com/balkey/eu1420/blob/master/makefile#L80

PostgreSQL limits the number of columns at 16 000 for a single table. Unfortunately, one of the Dutch files - endpoint belonging to programme code **2014NL16RFOP004** - contains more than 16 000 columns. These columns are declared in the .csv file (that is the format the source file is provided in), but contain no value. The file is clearly damaged, but this use-case called for the need to throw away all columns, which have all NULL values across all rows / in other words, we will discard a column if it is declared (even if it has a valid column name), but has no value for any of the rows within the file.

RENAMING FILES ACCORDING TO GENERATED PROGRAMME CODES AND PREPARING TABLE NAMES

https://github.com/balkey/eu1420/blob/master/makefile#L90

Finally we rename all files to their respective programme code, with a .csv extension. These files will still reside in a folder structure following the programme codes. If there are multiple files belonging to a programme code - either because there are multiple worksheets within a single endpoint containing a workbook, or there are multiple endpoints belonging to the same programme code - we sort the original filenames in alphabetical order (where the worksheet index maintains the original order of tabs) and assign an index to them delimited by an underscore. This will provide us with the following folder structure:

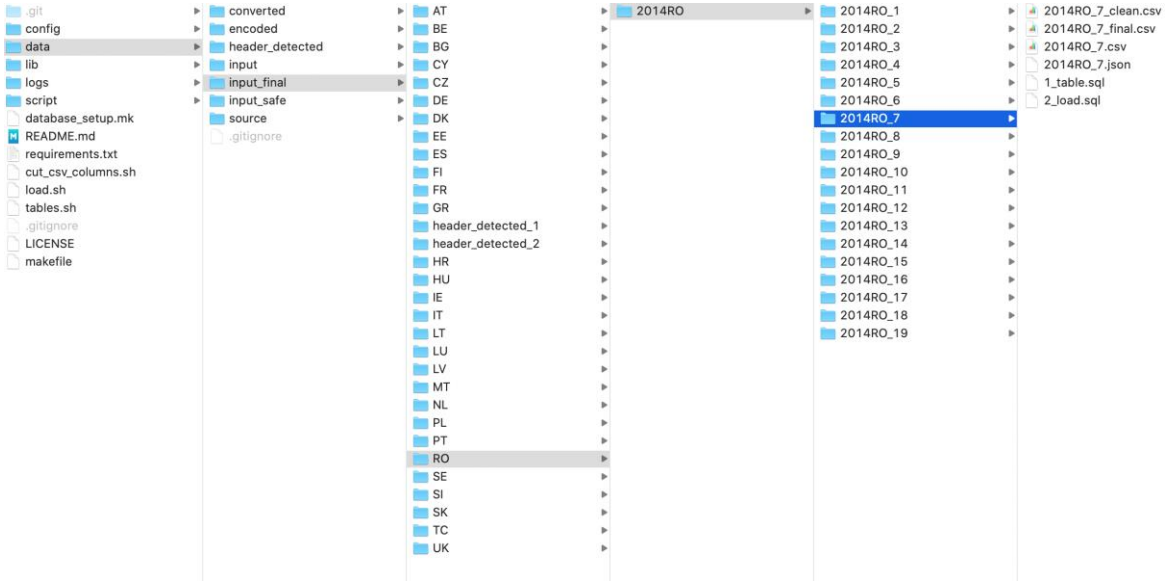

Figure 3.: folder structure autogenerated and ready for loading into database

This way we ensure a hierarchy that contains single files at the lowest branch of the tree, while maintaining Programme code relations to the fullest extent possible. The table names in the relational database will follow this convention, therefore there will be a table **named 2014RO_7** which contains the example file cited above.

AUTOGENERATION OF SQL TABLE CREATION SCRIPTS

https://github.com/balkey/eu1420/blob/master/makefile#L79

Once we have the above folder structure autogenerated and the corresponding validated .csv files, we can start autogenerating the table creation scripts for the database.

Since at this point we do not know what datatypes the specific columns contain, and we cannot be sure if the values themselves are actually valid to those datatypes, we will create a raw schema and materialise the table within that schema. As we want a 1:1 representation of the input file (which at this point includes the above mentioned transformations), we will just declare everything as text, and later assign strict datatypes to each column as the data cleaning and column value transformations take place.

The column names are declared after the transformed and validated header files, order of columns in the original source file is also maintained.

As mentioned earlier, the tables will be named after the programme codes maintaining uniqueness, and we also ensure that the job will run even if the database already has the specified table created (by always dropping the table first and then recreating it with each job / load).

Example of an autogenerated table creation script:

```
      DROP TABLE IF EXISTS raw."2014HR16M1OP001_1";

      CREATE TABLE raw."2014HR16M1OP001_1" (
        "naziv_projekta" TEXT COLLATE "default",
        "fond" TEXT COLLATE "default",
        "operativni_program" TEXT COLLATE "default",
        "nutsii" TEXT COLLATE "default",
        "zupanija" TEXT COLLATE "default",
        "korisnik" TEXT COLLATE "default",
        "opis_projekta" TEXT COLLATE "default",
        "datum_ugovaranja" TEXT COLLATE "default",
        "zakljucni_datum_provedbe_aktivnosti" TEXT COLLATE "default",
        "bespovratna_sredstva" TEXT COLLATE "default",
        "ukupno_prihvatljivi_troskovi" TEXT COLLATE "default",
        "status_projekta" TEXT COLLATE "default"
      )

      WITH(OIDS=false);
]
```

AUTOGENERATION OF SQL LOADING SCRIPTS

https://github.com/balkey/eu1420/blob/master/makefile#L79

Once we have the table creation scripts autogenerated, we just need to autogenerate the loading scripts. This is very similar to the table creation script, with one important detail: by declaring the column names themselves in the loading script, we ensure that only these columns will be loaded.

If the source file contains some rogue rows or shifted columns, where actually there's more values than the declared columns (a common problem with the .csv file format), we will bypass this as we declare the exact columns we would want to load and those are the exact same columns we created the table with (source of both is the validated header).

This ensures that the loading will happen without errors.

Example of autogenerated load script:

```
      \copy raw."2014HR16M1OP001_1" (
        "naziv_projekta",
        "fond",
        "operativni_program",
        "nutsii",
        "zupanija",
        "korisnik",
        "opis_projekta",
        "datum_ugovaranja",
        "zakljucni_datum_provedbe_aktivnosti",
        "bespovratna_sredstva",
```

```
  "ukupno_prihvatljivi_troskovi",
  "status_projekta")
FROM 'data/input_final/HR/2014HR16M1OP001/2014HR16M1OP001_1/
2014HR16M1OP001_1_final.csv'
DELIMITER ','
CSV HEADER;
```

PREPARING DATABASE AND AUTO-CREATING SCHEMAS

https://github.com/balkey/eu1420/blob/master/makefile#L126

We prepare the database in a standalone manner, meaning that with each run every dependency (schema creation, creation of user defined functions, granting user privileges, building indexes if necessary, etc.) is built from scratch.

This way we ensure that the project can be deployed at any newly created database instance, given that a user and a default database is created.

AUTOLOADING FILES INTO AUTOGENERATED SQL TABLES

https://github.com/balkey/eu1420/blob/master/makefile#L68

We automate the generation of tables with the help of the formerly described create scripts: the last input stage folder structure is walked through programatically, and if we find a file named **1_table.sql** we automatically execute it with parameterised variables (host, database, user, password, etc.) declared from configuration files.

After creating all the necessary tables, we load the files into the tables with the detection of **2_load.sql** scripts in the same manner.

Errors are piped into separate log-files dedicated to this step.

COLUMN MAPPINGS, DATA TRANSFORMATIONS AND CLEANING

https://github.com/balkey/eu1420/blob/master/makefile#L65

In this step each raw table's fields are mapped to the final master database table's corresponding columns. Data type transformations are carried in the scope of a single table - so each table can have its unique transformation rules. The transformations include:

- **Date casting:** all native and locale date formats are universally casted to YYYY-MM-DD date formats. In case the date was invalid or not recognisable - for example 31/02/2017 - the field is left empty.
- **Numeric casting:** separator characters are excluded, and all values are stored as numerical values, without specifying scale or precision.
- **Text casting:** all string values are stored as text without specifying the length of the string.

Transformation scripts are collected by country, so eventually each country has its own table, containing all collected transactions within the respective country.

CURRENCY CONVERSIONS

https://github.com/balkey/eu1420/blob/master/makefile#L61

For those countries which present their data in currencies other than EUR, a currency conversion is carried out, providing all values in EUR as well. For the exchange the 2019 yearly average rate is used, provided by Eurostat here.

MATERIALISATION OF THE MASTER TABLE

https://github.com/balkey/eu1420/blob/master/makefile#L62

In this final step all country specific tables are merged into one master table. One additional rule is included: all those rows, which have missing values for both **Operation name** and **Beneficiary name** are excluded from the final dataset.

EXPORTING THE MASTER TABLE

https://github.com/balkey/eu1420/blob/master/makefile#L58

The pipeline automatically generates a .csv export and saves it within the **data/exports/ folder.**

WORKFLOW MANAGEMENT

The above detailed workflow is orchestrated by makefiles, Python and shell scripts. Workflow dependencies are declared in the makefiles.

Environmental variables are stored and loaded from configuration files.

VERSION CONTROL

The project's source code is stored in a Github repository, which is currently a private repository aligned with the Commission's privacy requirements (the source code at this point should not be publicly accessible).

Upon request access can be provided to the repository EU1420.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

– at the following standard number: +32 22999696 or

– by email via: https://europa.eu/european-union/contact_en

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

KN-04-20-616-EN-N

ISBN: 978-92-76-23871-3