Cross-border data collection for journalists in Europe

EUROPEAN
DATA JOURNALISM
NETWORK

divergente

The following guidelines are based on the lessons learnt from ***The non-voter time bomb***, a data-driven cross-border journalistic investigation carried out in the 27 countries of the European Union (EU). All examples are taken from the European context.

Start with something small and build from there

Don’t put too much pressure on yourself to undertake a massive project but try to use a bit of data in any story you want to tell. Then, every time you want to cover a new topic, remember data can be your friend, something that will help you to tell a better story: use what you have already learnt from previous experiences and, where possible, build more on top of that.

This kind of practice, where you repeat little and often, is what really helps you to improve over time.

Write a synopsis of the project you are trying to create

This will help you to clarify your ideas and to convince partners that the project is worth their time and effort! Explaining your idea in one paragraph shows that you know exactly what your project is about.

Don’t forget to include a project briefing with:

- leading questions
- local/regional/national/transnational relevance
- level of the data (NUTS I, II or III)
- risks and constraints
- timeline
- contact person (in charge of coordination)

Schedule a kick-off meeting with partners

- Present and discuss the data unit (the subject of the data-driven project) based on the project synopsis you shared prior to the meeting.
- Leave some time for sharing: questions, doubts and concerns.
- Ask for partners’ input and thoughts and add them to your project.

Define the scope of the data-driven project according to the division of European regions

- If your project covers the majority of EU countries or focuses on a specific group of countries, it will be useful to collect data by dividing countries into northern, southern, eastern and/or western region(s).
- Data should also be presented according to these categories.

Determine the level of the data to be collected using a common scale

- NUTS—Nomenclature of territorial units for statistics

Be careful! The more disaggregated the data, the harder it is to compare (but it's not impossible!).

Much data is already available at NUTS I and II level, in open and reliable databases such as Eurostat. But remember: for fact-checking, you should use the primary sources, i.e. compare Eurostat datasets with the national source they come from.

Establish a partnership with academia for quantitative and qualitative data analysis

Data scientists are properly trained to handle numbers efficiently and reliably, they are used to searching for data in official datasets and know what is available and what is not. This makes them essential for defining the most feasible indicators to be collected in a group of countries for the subject under analysis.

List the official sources to be used by all partners

Be specific!

Include a brief description of the bureaus and offices you have in mind—different countries may have different names for the same responsibilities.

Example:

In Greece, the Ministry of the Interior is the bureau in charge of collecting and publishing all election results; in Portugal, it is the Electoral Commission who is responsible for collecting and publishing the election results at NUTS I and II level, and the Ministry of the Interior is responsible for data at NUTS III level.

List the official sources to be used by all partners

Be specific!

1. Add a short description including important information to ensure comparability.

Examples:

Blank vote: the ballot paper does not contain any mark or sign.

Null vote:

- more than one square has been marked on the ballot paper
- there is doubt as to which square has been ticked
- the square corresponding to a rejected or withdrawn candidate has been marked
- any cutting, drawing or erasing has been done
- any word has been written

List the official sources to be used by all partners

Be specific!

Illiteracy: People aged 10 and over that cannot read or write in a given reference period. Notes: The age above which a person attending the normal education system should know how to read and write is considered to be 10 years, equivalent to completion of the first stage of basic education.

University degree: Total number of individuals who have completed any tertiary education degree.

Unemployed: People aged between 16 and 74 who, during the reference period, met all of the following conditions simultaneously: 1) did not have a job nor was in work; 2) had actively sought work, i.e. had actively searched for a paid or unpaid job during the specified period (reference period or the three previous weeks); 3) was available for a paid or unpaid job.

Notes: It includes people who have already found a job but who would only start working on a date after the reference period (the following three months).

List the official sources to be used by all partners

Be specific!

2. Take advantage of the European standards and classifications.

Examples:

NACE Code—Nomenclature of Economic Activities for information about business activities.

ISCED—International Standard Classification of Education for information about levels of education among the population.

Data scientists have extensive knowledge of these classifications. They can help you understand the standard classifications and which ones are best suited to your research.

✓ The list of concepts must be shared with all partners from the outset of the research, to identify which can be compared and to decide on the final list of data to be collected.

✕ Don't ask for unnecessary information. Put yourself in the other person's shoes: everyone has a lot going on and no one likes to feel like they're wasting their time.

Prepare a step-by-step document to be followed throughout the data collection process

Don't forget to include:

- an editable spreadsheet for each partner with pre-agreed indicators but also with the basic information needed to create filters during the data analysis, i.e. country acronym, country names, latitude and longitude for interactive mapping, European region and year the data refers to
- a shared document for partners to note down adjustments and exceptions needed during the data collection

Prepare a step-by-step document to be followed throughout the data collection process

Don't forget to include:

- dots or commas between decimal places
- cells formatted for % or entered manually
- nomenclature of categories in a clear and uniform way (e.g. the names of elections—National or Parliamentary)
- the language to be used in the original language or in English (e.g. name of parishes)

This will avoid future problems when using filters, formulas and analysing the data.

Prepare a step-by-step document to be followed throughout the data collection process

✓ Set deadlines for updates on the data collection process and final delivery.

✓ Ask for links to the original source for each piece of data, so that you can quickly confirm the sources.

✓ Ask for official documents to be uploaded to a folder, preferably in machine-readable formats such as .xlsx or .csv. These are essential for the fact-checking phase.

Be realistic on timing. Data collection depends not only on the efforts of the journalist but also on how easy or difficult it is to access the data, which will impact the time needed to complete the task.

Be available to clarify partners' doubts during the data collection

And always make sure that the data being collected meets the following requirements:

- accuracy according to standard data
- consistency with other data sets (from other partners' sources)
- harmonization by using the same units of measurement

Making partners wait too long for a response or update can jeopardise the efficiency of the data collection and the commitment and availability of partners to continue the task.

Save resources and time when reviewing data

Sometimes the original datasets are only available in formats other than .xlsx or .csv and must be manually entered. This process has a higher margin of error, and not every newsroom has a data team. So it's likely that data collection will be done by people who don't just work with data on a daily basis.

✓ Double-check ALL data collected by the partners

✓ Combine all data into a MASTER_DOC to be shared with all partners and statistical experts for data analysis

Ask questions of the data

Ask the data scientists to apply statistical calculations—Spearman's coefficient (ρho) or Pearson’s coefficient (r), for example—to highlight the outliers and the trends in the numbers.

Ask questions of the data

✓ Only base the journalistic approach on correlations classified as moderate to very strong.

✕ Don't confuse correlation with causation. Adapting an analytical methodology to arrive at trends is different from identifying the direct causes of them.

✕ Don't consider the correlations classified as weak for the journalistic approach.

Don't forget your initial questions and the scope of the research. But also keep an open mind to what stories the numbers are telling you (particularly if it is something that you had not thought of to start with).

Write a methodology paper to be shared with all partners

This document is based on the data scientist’s report and the conclusions of the quantitative and qualitative analysis. It is useful for partners to know what they can rely on, the choices made by the statistical experts during the analysis, and will also be available on the website of the feature.

See ***The non-voter time bomb* methodology** as an example.

Data collection as a starting point

Data says everything and its opposite! So, don't blindly trust data without treating it with caution.

And more importantly: **journalistic work must be based on a variety of sources. Data is only one of them.**

So, discover the stories hidden within the numbers and go find the news they tell. **Good luck!**

EUROPEAN
DATA JOURNALISM
NETWORK

divergente