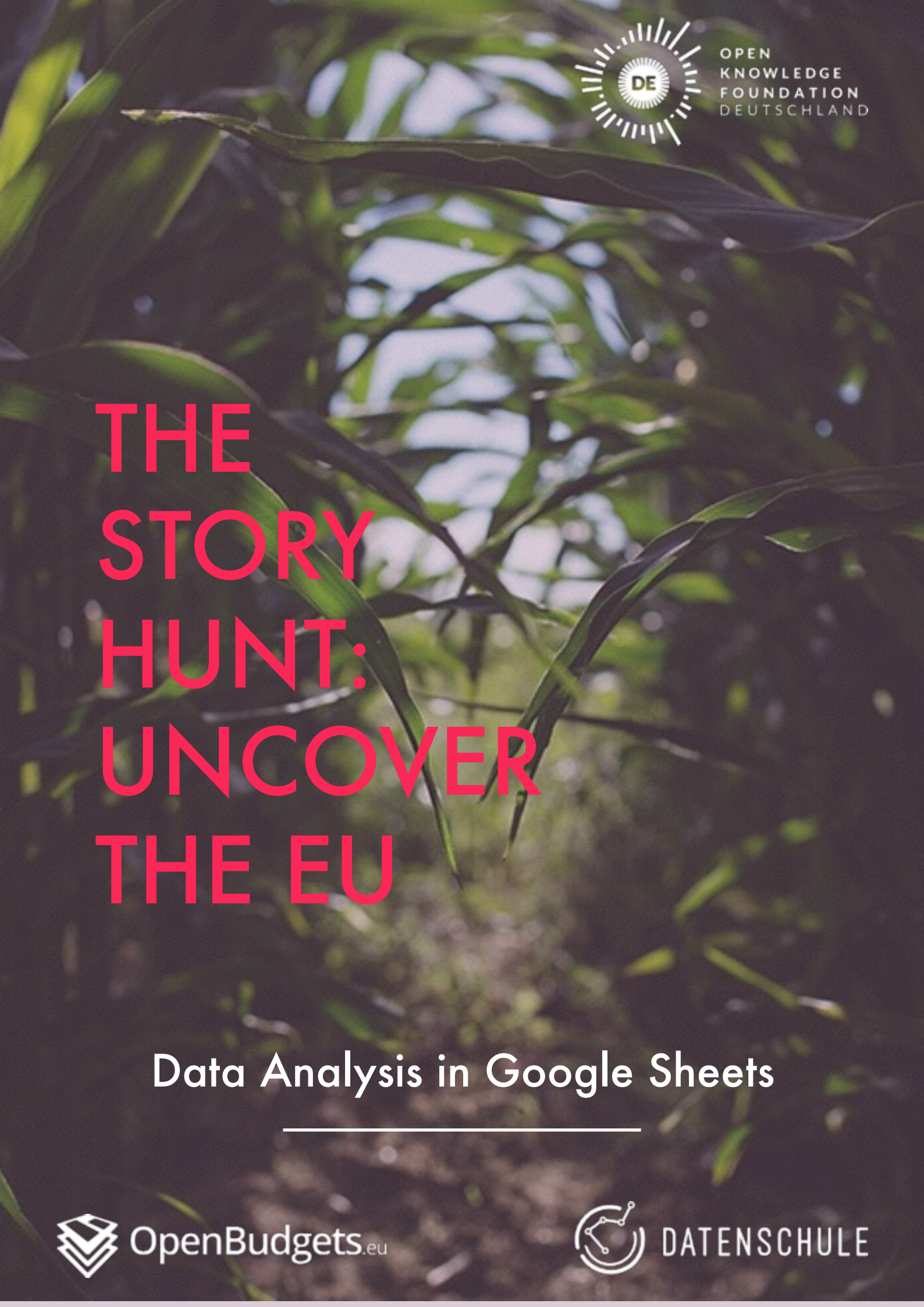

OPEN
KNOWLEDGE
FOUNDATION
DEUTSCHLAND
THE STORY HUNT: UNCOVER THE EU
Data Analysis in Google Sheets
OpenBudgets.eu
DATENSCHULE

Data-Analysis in Google- Sheets

By analyzing Data we obtain more insights into a given question, test our hypothesis or obtain new leads for further research. There is an abundance of different data analysis methodologies ranging from descriptive and inferential statistics to probability stats. This learning material provides general approaches when analyzing a dataset as well as links to useful programs and resources which might prove helpful for your project.

Table of Content

Data Analysis

Getting a feel for the data:

- When beginning a Data-analysis it is vital to first get a feel for the information within the dataset. You might ask yourself the following Questions:
  - How many observations are in my dataset (number of rows), And how many columns are included ?
  - Does the dataset contain time series data ?

- *Next it is helpful to search for:*
  - General Trends in Data (see: measures of central tendency)
  - Outliers, which values are particularly high/low; occur most often or the least?

- Context information is important too:
  - Who produced the dataset?
  - When was it produced?
  - What is the intention of it?

These Questions are helpful for verifying the data.

Further questions might be:

- How are columns related?
- Can we relate the main findings to the topic of the dataset?
- Do our findings prove or disprove established facts or theories?

Measures of central tendency

Measures of Central Tendency:
Measures of Central Tendency are summary statistics which aim to describe a set of data in one specific number. The main stats are described in the following

Mean: The Average:
- What is the average EU subsidy a member state receives? =AVERAGE(A1:A29)
- All data points divided by the number of observations

Median: the value in the middle:
- the value that separates the dataset into two equal halves (50 / 50)
- What value lies exactly in the middle of the distribution? =MEDIAN(A1:A29)
- Very useful for income datasets - helps negate the high income bias that might be introduced through

Mode: most frequent value:
- datapoint that appears the most in your dataset =MODE(A1:A29)
- This only makes sense with absolute numbers or low amount of decimal places
- For example school grades: which is the most frequent grade?

Measures of central tendency

Measure of Spread:
Minimum: lowest value in dataset =MIN(A1:A29)
Maximum: highest value in dataset =MAX(A1:A29)

Standard deviation: Measures what is "normal" or expected

- "standard deviation is the average distance to the average" =STDEV(A1:A29)
- the standard deviation tells you by how much any given value within the dataset is expected to spread from the mean

Correlations

Correlation:

- A correlation is a statistical relationship between two variables
- For example: We observe a high amount of storks and babies being born in the same region!
- Does this imply that the stork delivers the babies? (a "causal" relationship?)
- Our common sense tells us NO!
- We forgot to account for the variable region → there is a higher stork population in the countryside and there are more babies being born in the countryside
- "Correlation does not imply causation!"

When is causation appropriate?

- Follow your common sense, if a relationship is too good to be true, be critical
- Only "believe" a causal statement if you cannot think of any missing information

Example for a causal relationship:

- Correlation between chicken wing sales on Sundays
- A large spike on the first Sunday in February
- Can the spike in wing sales be plausibly explained by the Super Bowl taking place?
- Yes! At least we do not know any other reason, why wing sales would increase so drastically!

EU-Financial Data

Analyzing financial data can be hard to handle, due to its complexity!

- When looking at Budgets it is crucial:
  - to understand the nature of the data → context
  - Where does this data come from?
  - What do the numbers represent?
  - What time period does it include?

How can you relate large numbers?

- The amount of EU subsidies received per country is often in the billions
- Is there a way to put the number into perspective? Relation to other large numbers!
- How much of the national gross domestic product does the subsidy account for?
- Beware only compare equal to equal (subsidies are paid over 7 years)

Links & Programs

Online Resources:

1. The DataScience Academy has an extensive list of free resources to get started.
    1. http://datascienceacademy.com/free-data-science-courses/
2. DataCamp offers many Data Science Courses for Python, R and SQL. Basic ones are usually free:
    1. https://www.datacamp.com/
3. Guess the Correlation: http://guessthecorrelation.com/
    A fun game for introducing the concept of correlation

Tools:

4. Stata http://www.stata.com/:
    1. Probably the most used statistical analysis tool in the field of academics, yet the tool is not open source.
5. R https://www.r-project.org/ & R Studio https://www.rstudio.com/
    1. Another very popular programming language for statistical analysis. The tool R-Studio provides an interface for the analysis. The tool is open source
6. Python Libraries:
    1. The Programming language python also provides very useful libraries for statistical analysis such as pandas, numpy and scikit. The ipython-notebook is a good user-interface for this. It helps you to structure your analysis.

About Us

OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND

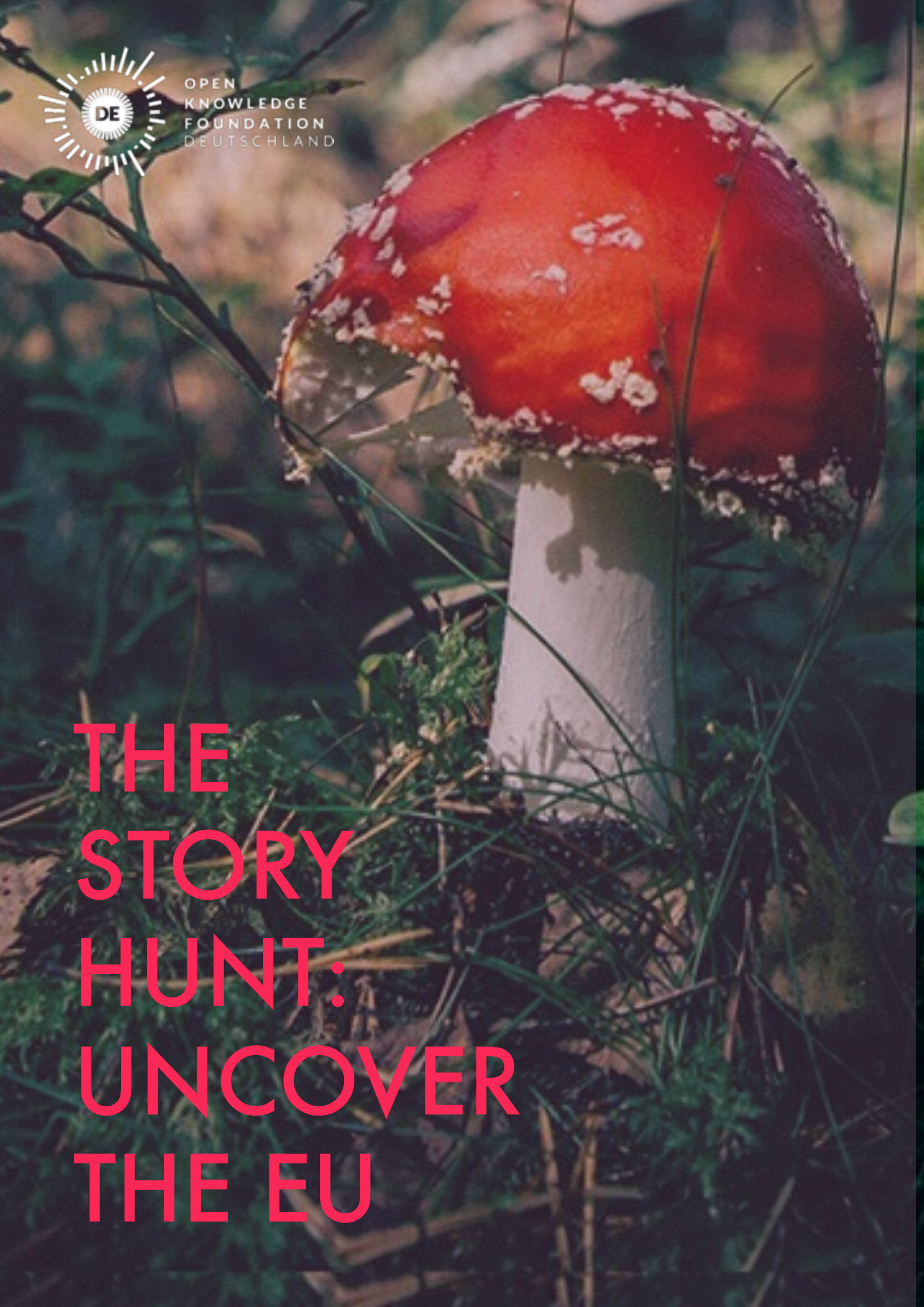
The Open Knowledge Foundation Germany is a nonprofit organization that advocates open knowledge, open data, transparency, and civil participation.

OpenBudgets.eu

OpenBudgets is an EU funded project, aiming at supporting journalists, civil society organisations, NGOs, citizens and public administrations, by providing an overview of public spending, as well as tools and appropriate data and stories to advocate and fight for fiscal transparency.

DATENSCHULE

School of Data Germany helps non-profit organisations, civil rights defenders and activists to understand and use data and technology effectively to increase their impact on societal challenges.


OPEN
KNOWLEDGE
FOUNDATION
DEUTSCHLAND
THE
STORY
HUNT:
UNCOVER
THE EU