BEGINNGERS GUIDE

DATA ACCESS & EXPLORATION

Data Access and Exploration: A beginner's guide

Understanding and utilising data enables individuals to make informed decisions, leading to more effective policies and projects that significantly enhance people's lives. Data literacy has never been a more essential skill for those working in public administration, academia, and the public sector.

The concept of "data for good" embodies the potential of data to drive positive social impact. For instance, public health data analysis can track disease outbreaks, allocate resources more efficiently, and develop preventive measures. In environmental protection, data helps monitor pollution levels, assess the effectiveness of sustainability initiatives, and guide policy decisions that safeguard natural resources and public health. In mobility, data can optimise public transportation systems, reduce traffic congestion, and improve urban planning. In green transformation, data helps monitor renewable energy production, assess the impact of green technologies, and guide investments in sustainable infrastructure.

Accessing and analysing data empowers researchers and policymakers to uncover insights that would otherwise remain hidden. This capability leads to better-informed decisions addressing societal challenges more effectively.

Building on the insights from our recent webinar on "Data Access and Exploration", this guide expands on key areas and provides additional resources for beginners. Whether you want to enhance your data literacy or start learning data analysis, this guide will help you grasp the fundamentals.

What this guide will cover:

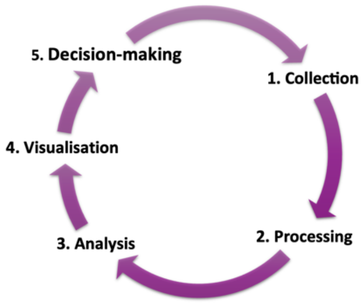
1. The Importance of Data Literacy: How data literacy empowers public administration and academia to make better decisions.
2. Understanding Data Access: Different types of data, as well as legal and ethical considerations in accessing and using data.
3. Introduction to the Big Data Test Infrastructure (BDTI): Overview of BDTI and the tools and resources provided by BDTI to support data projects.
4. Step-by-Step Guide to Data Access: Identifying your data needs, finding reliable sources, evaluating data quality and preparing data for analysis.
5. Exploring Data: Tools and techniques for data exploration. Initial steps to understand your datasets and uncover insights.
6. Use Case: Investigating Funding of Green Energy Projects in EU Universities
7. Practical Exercises and Further Learning

Let's dive in.

1. The importance of data literacy

The demand for professionals with data analysis skills is experiencing exponential growth due to the need for data-informed decisions across industries. The European Year of Skills initiative underscores the importance of these competencies, particularly as the digital and green transitions create new opportunities and challenges. For public administration and academia, data literacy means improved efficiency, evidence-based policy-making, and enhanced research capabilities.

Becoming au-fait with data begins with understanding the data journey. When we know the full-cycle impact data has, we can properly grasp its power.



What is the data journey?

The data journey refers to the various stages by which data moves from collection to use. The journey involves several key steps:

1. Collection	Define	Defining the purpose and scope of the data collection
	Find	Identifying the sources of the data required
	Gather	Collecting and aggregating the data the identified sources
2. Processing	Explore	Exploring the data to understand its quality, completeness, and relevance
	Clean	Cleaning and preprocessing the data to ensure its accuracy and consistency
	Describe	Describing the data, including its structure, format, and content (optional, but helpful)
3. Analyse	Analyse	Analysing the data to extract insights and meaning
	Model	Modeling the data to make predictions, identify patterns, and forecast future trends
4. Visualise	Storytell	Presenting the findings and insights from the data in a clear and meaningful way
5. Decision making		

We'll discuss *define*, *find*, *gather* and *explore* in this guide, and the rest of the journey in future guides.

2. Understanding data access

Data access refers to locating, retrieving, and using data effectively. It is the critical first step in any data analysis project. And at this point, it's helpful to understand the different types of data, including open data, proprietary data, and personal data, each with its implications and uses.

Types of Data:

Open data is freely available data that can be used and shared without restrictions. For example, the datasets available on data.europa.eu.

Proprietary data is owned by an individual or organisation and typically requires permission or a license.

Personal data is information about an identifiable individual, which must be handled with care due to privacy concerns.

Legal and Ethical Considerations: Accessing data, especially personal and proprietary data, involves legal and ethical considerations. The EU's General Data Protection Regulation (GDPR) sets strict guidelines on how personal data should be handled to protect individuals' privacy. When working with data, it's crucial to ensure compliance with legal standards and ethical practices, such as obtaining necessary permissions and anonymising personal information to protect privacy. European acts such as the Data Act and the Open Data Directive are cornerstones of the union's drive towards open data access, use and reuse. The Big Data Test Infrastructure (BDTI) is a key enabler of these drives.

3. Introduction to the Big Data Test Infrastructure

The Big Data Test Infrastructure (BDTI) is an initiative funded by the Digital Europe Programme (DEP) designed to support public administrations in acquiring and building the skills to extract value from data. BDTI offers a comprehensive suite of open-source tools and a cloud-based platform to facilitate data projects leading to data-informed decisions. Public administrations can apply for a BDTI pilot project, which provides six months of free access to these resources.

Tools and resources provided by BDTI:

BDTI includes tools supporting various data analysis stages, from collection to visualisation. In this guide, we'll discuss those used during the data access stage and introduce tools to perform subsequent steps in future guides.

For data access and exploration, we'll see three tools in action: KNIME, known for its user-friendly interface and visual workflows; R-Studio, a powerful tool for statistical analysis; and Jupyter Notebooks using Python, an easy-to-use interactive data exploration. These are powerful tools yet very accessible for beginners to grasp and use.

4. Data Access: A step-by-step guide

Finding and accessing relevant datasets is a crucial skill for any data analyst. This stage typically covers the first four steps outlined in the table above and others:

Identify your data needs

Identifying your data needs is a crucial first step, and it involves understanding what specific information you require to achieve your goals. Here are some key considerations and questions to ask yourself to help identify your data needs:

Key Considerations

Objective: What is the primary goal of your project? Are you looking to solve a specific problem, answer a question, or test a hypothesis?

Scope: What is the geographical scope of your analysis (local, regional, national, or global)? What is the time frame for the data (historical data, current data, future projections)?

Specificity: What level of detail is required (granular data at the individual level, aggregated data at the regional level)? Do you need quantitative data (numbers, statistics) or qualitative data (descriptions, opinions)?

Sources: What types of data sources are relevant (open data portals, proprietary databases, surveys, sensors)? Can existing datasets be reused, or do you need to collect new data? What legal and ethical due diligence is required?

Format and structure: In what format should the data be (CSV, JSON, Excel, SQL databases)? What structure should the data follow (time-series, relational, geospatial data)?

Quality and reliability: What standards must the data meet (accuracy, completeness, timeliness)? How will you verify the reliability and validity of the data sources?

Are the different datasets interoperable? Data **interoperability** refers to the ways in which data is formatted that allow diverse **datasets** to be merged or aggregated in meaningful ways. It is a key aspect of the FAIR Data Principles, constituting the "I" in FAIR. Interoperable datasets are a crucial aspect of the European Union's data strategy, aiming to make publicly funded information available for new information products and innovation, particularly in artificial intelligence. The EU policy framework on data interoperability is articulated, with the Open Data Directive playing a key role in promoting data sharing and reuse.

Use reliable sources

Ensuring the reliability of your data sources is crucial for the credibility and accuracy of your analysis. Here are some key considerations and prerequisites for determining the reliability of a data source:

Key Considerations

Source Reputation: Consider the authority of the data provider. Reliable sources are often well-known organisations, government bodies, academic institutions, or reputable research organisations. Look for sources with a history of providing accurate and high-quality data. Search for data on reputable open data portals such as data.europa.eu and Our World in Data. These platforms offer a vast array of freely available and well-documented datasets.

Data Documentation: Reliable data sources provide comprehensive metadata, including information on how the data was collected, processed, and validated.

Transparent documentation of the data collection methodology helps assess the reliability and relevance of the data.

Consistency and Timeliness: Reliable data sources regularly update their datasets to reflect current information. Data that is consistent over time and across different datasets indicates a reliable source.

Community Feedback: Sources that are widely used and reviewed by the data community can provide insights into the reliability and usability of the data. Platforms that allow users to rate and review datasets can help identify reliable sources.

Where can I find open datasets?

Government Data Portals

- Member states each have national data portals like data.gov.ie, data.gov.gr, etc.
- Member state portals are also accessible from data.europa.eu, The official portal for European data

- Specialised Data Repositories like;
  - European Environment Agency (EEA): Offers datasets related to environmental issues across Europe.
  - Eurostat: The statistical office of the European Union provides data on various economic and social aspects of the EU.

International Organisations, for example;

- World Bank Open Data: Offers extensive datasets on global development indicators.
- United Nations Data: Provides access to datasets from various UN agencies, covering a wide range of global issues.
- OECD Data: The Organisation for Economic Co-operation and Development provides data on economic indicators and social statistics.

Research Institutions such as;

EUI Research Repository: Offers access to a variety of open datasets and research outputs related to social sciences, economics, and political science.

ESS Data: Provides open access to high-quality data on attitudes, beliefs, and behaviour patterns of various populations in Europe.

Academia and universities:

- Amsterdam Open Research: A platform for sharing open research data and publications, promoting open access in various fields.
- Tilburg University Data Repository: Offers open access to datasets and research outputs focusing on social sciences, law, and economics.
- AUSSDA – The Austrian Social Science Data Archive: Provides open access to social science datasets, supporting research and teaching.

Non-Governmental Organisations (NGOs) and Non-Profits like;

- Our World in Data: Provides datasets on global challenges such as health, environment, and social issues.

Evaluate data quality

By this point, much of the work that goes into evaluating data quality has been done, but it's worth reiterating the importance of quality when it comes to data analysis. Findings are only ever as useful as the data used to get there.

Why evaluate my data?

Data quality is paramount in any data analysis project as it directly impacts your findings' accuracy, reliability, and validity. High-quality data ensures that insights are trustworthy, and decisions based on this data are sound.

To ensure data quality, think CAR. 🚗

Completeness	**Accuracy**	**Relevance:**
Ensure all necessary data fields are filled, and there are no missing values.	Verify the data against other reliable sources to confirm its correctness.	Ensure the data is pertinent to your research question or project objectives.

Start by assessing the accuracy of the data through cross-verification with reliable sources. Check for completeness by identifying and addressing any missing values. Consistency can be maintained by ensuring that related data fields do not contain conflicting information and that relationships between datasets are intact. Regularly update your data to maintain timeliness and verify that it conforms to expected formats and business rules to ensure validity. \

By focusing on these aspects, you can maintain high data quality, leading to more effective and credible analysis.

Download and prepare data:

Once you've found suitable datasets, it's time to download them in accessible formats and 'read in' the data.

What does 'read in' mean?

'Reading in' data refers to the process of loading data from a file or external source into an environment or application for analysis and manipulation. It's a step in the data analysis workflows as it transforms raw data into a format that can be understood by analytical tools.

Common tasks involved in reading in data include specifying the file path, selecting the appropriate file type, and using specific functions or commands to import the data into the working environment. For example, in Python, the pandas library is often used to read in data from CSV files using the read_csv() function. Similarly, in R, the read.csv() function serves the same purpose. This process ensures the data is accessible and ready for further analysis and exploration.

Reading in data resources:

Read in data using Python

Loading data into R

Common data file format: a brief overview

Some of the most used data file formats are CSV, JSON, XML and Excel. Here is a brief primer on each of them and what they are most used for.

CSV (Comma-Separated Values)

Description: A plain text format that uses commas to separate values. Each line in the file corresponds to a row in a table, with columns separated by commas.

Use Cases: Ideal for storing tabular data, widely supported by spreadsheet applications and data analysis tools.

JSON (JavaScript Object Notation)

Description: A lightweight data-interchange format that uses human-readable text to store and transmit data objects consisting of key-value pairs and arrays.

Use Cases: Commonly used for APIs and web applications due to its simplicity and ease of use with programming languages.

XML (Extensible Markup Language)

Description: A flexible, structured format that uses tags to define objects and data within a hierarchy.

Use Cases: Often used for data exchange between systems and applications, especially where a rigorous data schema is needed.

Excel (XLS/XLSX)

Description: Proprietary spreadsheet formats used by Microsoft Excel, capable of storing data in multiple sheets along with complex formulas and formatting.

Use Cases: Commonly used in business environments for data analysis, reporting, and financial calculations.

4. Exploring data

Data exploration involves examining your datasets to understand their structure and identify patterns or insights. The initial steps involve data extraction and cleaning, such as removing duplicates and dealing with empty cells. Several open-source tools can help with this process, including R-Studio and Jupyter Notebook.

R-Studio: A powerful tool for statistical analysis and data visualisation. While it requires some knowledge of the R programming language, it offers extensive data manipulation and analysis capabilities. R-Studio provides a comprehensive environment for data analysis, including tools for importing data, cleaning it, and performing advanced statistical analyses. RStudio Documentation

Resources for data exploration using R-Studio:

R for Data Science

Jupyter Notebook: Jupyter Notebook supports several programming languages, including Python. It's excellent for interactive data analysis and allows you to create and share documents that contain live code, equations, visualisations, and narrative text. Jupyter Notebook is particularly useful for documenting your data analysis process and creating reproducible research.

Data exploration with Jupyter Notebook resources:

Python Data Science Handbook

KNIME Analytics Platform: KNIME is a user-friendly, open-source tool that allows for visual programming. It's ideal for beginners as it requires no coding skills. You can import datasets, perform basic data cleaning, and create workflows for data analysis. KNIME supports various data types and offers a wide range of built-in nodes for different data processing tasks.

Beginner KNIME resources:

KNIME beginners quick-start guide

Use case: Investigating funding of green energy projects in EU universities

During the recent webinar on Data Access, participants followed along with a use case involving the investigation of funding for green energy projects in EU universities. Let's briefly look at the case and the approach we took to illustrate the data access and exploration process.

1. **Define objectives**

- Visualise energy projects: Identify and map energy projects funded by Horizon Europe.
- Identify universities receiving funding: Determine which universities have received funding for green energy projects.
- Explore the relationship between CO2 emissions and project funding: Analyse if there is a correlation between the funding for energy projects and the CO2 emissions of the countries where the universities are located.

2. **Find relevant data**

- Project funding information: We searched for datasets on data.europa.eu using keywords like "Horizon 2020" and "Horizon 2021" to find information on EU-funded projects.
- CO2 emissions data: We accessed Our World in Data to find datasets related to CO2 emissions.

3. **Prepare and explore data**

- Data extraction: Downloaded datasets in CSV format from the sources above.
- Data cleaning: Initial steps to clean the datasets, such as removing duplicates, standardising formats, and ensuring consistency using tools like OpenRefine.

- Example: The Horizon 2020 dataset included columns such as project ID, project name, funding amount, and participating organisations. Filtered the data to focus only on energy-related projects.
- CO2 dataset: Filtered to include only EU member states and relevant emissions data.
- Basic exploration: Using KNIME, R-Studio, or Jupyter Notebook to perform basic exploration such as viewing summary statistics, identifying key variables, and getting an overview of the dataset structure.

Practical exercises and further learning

To reinforce your learning, use the hands-on exercises in the webinar's code repository. These exercises will help you practice data access and exploration techniques.

Download the tools:

- R-studio: https://cran.r-project.org/
- Jupyter Notebook: https://jupyter.org/install
- KNIME Analytic Platform: https://www.knime.com/downloads