#DigitalEUProgramme

Introduction to this document

This document provides guidance for public administrations interested in experimenting with the Big Data Test Infrastructure (BDTI).

The objective of this document is to provide **useful tools for users to define and design their pilot use cases** while leveraging the capabilities provided by BDTI.

Introduction to BDTI

BDTI was created by the **European Commission** in 2019 and is funded by the Digital Europe Programme (DEP) which aims to accelerate the economic recovery and shape the digital transformation of Europe's society and economy, increasing the **availability**, **quality** and **usability** of **public sector information** in compliance with the requirements of the Open Data Directive (2019).

The **DEP** is one of the activities supporting the implementation of the European Data Strategy, adopted by the European Commission in 2020, which aims to create a **single market for data**, ensuring that more **data become available for use in the economy and society in Europe**, keeping the companies and individuals who generate the data in control of them.

Introduction to data products

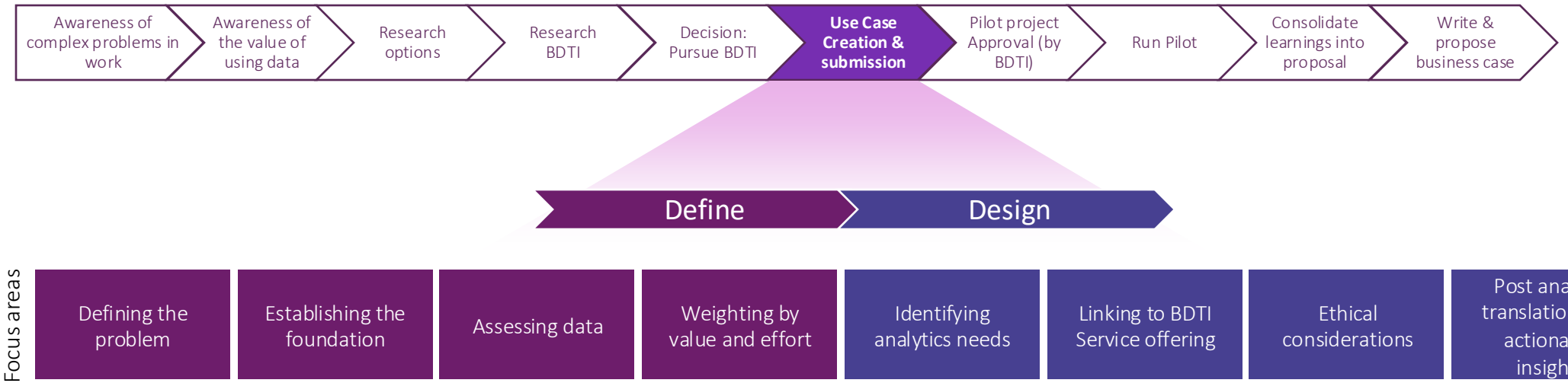
BDTI allows you to build a data product. **Data products** are specialised tools or applications designed to leverage data for specific purposes, providing **valuable insights, automation, or enhanced decision-making capabilities**.

These products can range from **simple data visualisations** and dashboards to complex machine learning models and **AI-driven applications**. They are built to process, analyse, and present data in a way that is easily **understandable** and **actionable** for users, helping the organisations to optimise operations, identify trends, and make informed policy decisions.

Data products are essential in today's data-driven world, enabling public administrations to **harness the power of their open data effectively**.

Methodology used in this document

Among all the actions to be considered when addressing a broader problem, this document focuses on the **Use Case Creation & Submission** process. It comprises two main sections, detailing the two phases of developing your use case – **Define** and **Design**, and further outlining their respective steps.



Inspiration is taken from the CRISP-ML(Q) Lifecycle Process, https://ml-ops.org/content/crisp-ml

Introduction to the document structure (1/3)

While browsing this document, you will encounter these **graphical cues** to help you navigate.

Theory
Describe the theoretical foundations for the analysis and tools.

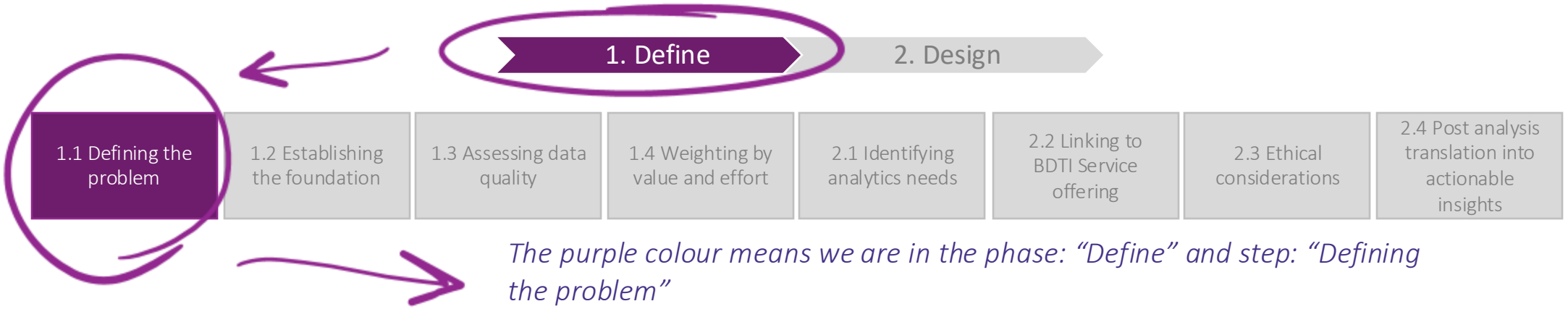
Tool
Facilitate the analysis, comprise useful questions, decision trees and other support for decision-making.

Important
Present important concepts to be considered more carefully.

Idea
Present ideas or examples for inspiration or better understanding of the concept.

Introduction to the document structure (2/3)

The beginning of each phase or step will be represented with the **introduction page**, highlighting with colour, **where you are** on the methodology roadmap.



The purple colour means we are in the phase: "Define" and step: "Defining the problem"

Additionally, on every page of this document, a visual cue in the bottom right corner is shown to **highlight the current phase**.

The purple colour means we are in the phase: "Define"

Introduction to the document structure (3/3)

This document is designed to support you in achieving your goals through a variety of tools, including checklists, decision trees, and frameworks.

Recognising that users are at **different stages of their data-driven journey**, the methodology is **flexible and adaptable**.

You have the **autonomy** to select the tools that are most relevant to your **current needs and maturity level**.

Begin by evaluating your position and requirements and then **choose the most appropriate starting point**.

You may not need to use all the tools provided; instead, **focus on those that will add the most value** to your specific situation.

Idea: Inspiration for data products

[Rooting for data: Revolutionising urban tree health monitoring in Bochum, Germany]

Building a machine learning model to revolutionise how tree health is monitored in urban areas.

[The MitosLOD Pilot Story: Transforming Greek public services with Linked Open Data]

GRNET and UoM aim to transform MITOS, which provides structured descriptions of over 3,000 public services, into Linked Open Data.

[Enabling Public Transport Efficiency: City of Turku, Finland]

A collaborative pilot involving the City of Turku and the University of Turku analysing mobility efficiency and public transportation using BDTI's free sandbox of tools.

[Mobility and Urban Environment Information for Decision Making: City of Napoli, Italy]

The Napoli pilot will test advanced analytics on data related to Public Space (parks, street areas, other) and mobility-related data to build decision support systems for evaluating potential future use of public space.

Discover more use cases on a dedicated BDTI website or in the inventory of use case [BDTI - D01 Inventory of Use Cases_v4.00.docx]

IDEA

Phase 1: Define

Define | Design

Scoping your pilot can be tricky. Use this chapter to help you define and refine your key objectives.

Define refers to defining the problem, establishing the foundation for the pilot, performing a data quality assessment, and looking into the value and effort of the pilot use case.

You can jump to the desired page by clicking the links

Defining the problem statement

At a high level, the definition of the problem statement contains the following steps:

1. **Identify the problem**
   Clearly describe the issue you're facing. Be specific about what is wrong and why it needs to be addressed.
2. **Put the problem into context**
   Examine the background information that helps explain the problem. This might include the environment, stakeholders, and any relevant history.
3. **State the problem's impact**
   Explain the consequences of the problem. How does it affect people, processes, or systems? Quantify the impact if possible.
4. **Set the objectives**
   Outline what you hope to achieve by solving the problem. These should be specific, measurable, achievable, relevant, and time-bound (SMART).
5. **Outline the benefits**
   Create a list of benefits you are expecting to come out of a project. Ideally, they should eliminate the problem's impact.

Inspiration taken from https://www.sixsigma-institute.org/Six_Sigma_DMAIC_Process_Define_Phase_Six_Sigma_Project_Charter.php

Idea: Example of a problem statement

1	2	3	4	5
Problem	Context	Impact	Objectives	Benefits
The city's public transportation system is experiencing significant delays, leading to decreased ridership and public dissatisfaction.	Over the past year, the average delay time for buses and trains has increased from 5 minutes to 15 minutes. This issue has been exacerbated by outdated infrastructure, a lack of investment in maintenance, and an increase in the number of passengers. Key stakeholders include the city government, public transportation authorities, commuters, and local businesses.	The delays are causing frustration among commuters, leading to a 20% decrease in ridership. This decline in usage has resulted in reduced fare revenue, which further limits the budget for necessary improvements. Additionally, local businesses are affected as employees and customers face difficulties in reaching their destinations on time.	•Reduce the average delay time for buses and trains to under 5 minutes within the next 12 months. •Increase ridership by 15% within the next 18 months. •Improve customer satisfaction scores related to public transportation by 25% within the next year.	•Enhanced reliability of the public transportation system, leading to increased public trust and usage. •Higher fare revenue, which can be reinvested into further improvements and maintenance. •Improved punctuality for commuters, resulting in better productivity and satisfaction. •Positive economic impact on local businesses due to more reliable transportation for employees and customers.

Tool: Questionnaire for problem assessment

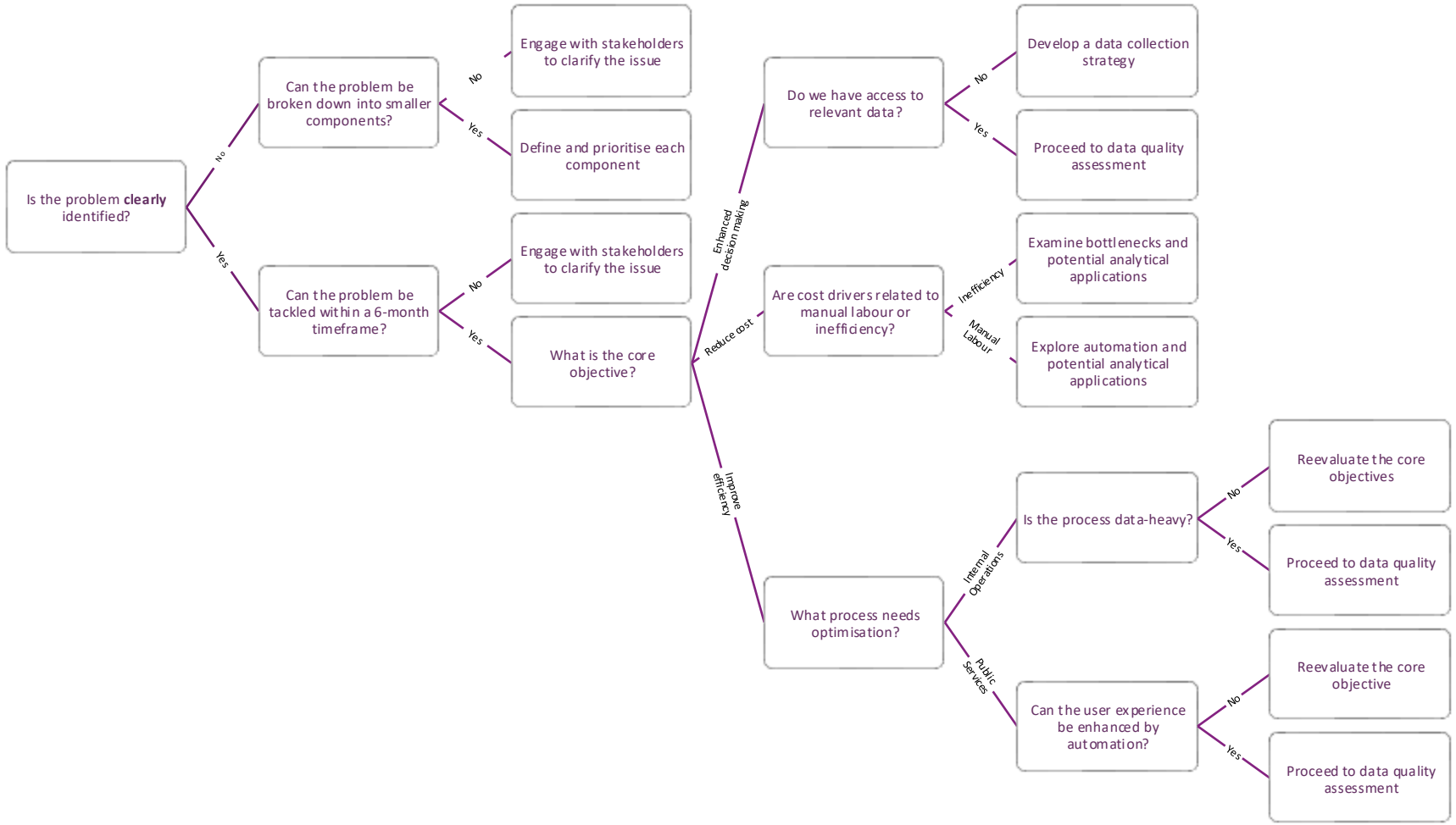
How do we assess the problem and identify the ideal outcome?

Topic	*Description*	*Questions*
Problem Statement	Clarify the Problem Statement A problem statement identifies a specific issue or challenge that affects the efficiency, effectiveness, or equity of public services. It succinctly outlines the nature of the problem, its impact on the community or stakeholders, and the need for a solution to improve public governance and service delivery.	❑ *What is the problem we are trying to address?* ❑ *What exact problem are we aiming to solve with data analytics?* ❑ *Who are the stakeholders or groups affected by this problem?* ❑ *What is the current impact or consequences of this problem on the community or organisation?* ❑ *What are the underlying causes or contributing factors to this problem?*
Objectives	Prioritise Objectives Objectives are clear, specific, and measurable targets set to achieve desired outcomes in public policies, programs, or initiatives. They provide direction and focus for actions, ensuring that efforts are aligned with broader goals of improving public welfare, efficiency, and service delivery. Objectives serve as benchmarks for evaluating progress and success in addressing public needs and challenges.	❑ *Which objectives are most critical to achieving the desired outcome?* ❑ *What are the specific outcomes or results we aim to achieve with this policy or program?* ❑ *Who are the primary beneficiaries or target groups of these objectives?* ❑ *What are the key actions or steps required to achieve these outcomes?* ❑ *How will we measure and evaluate the success of these objectives?* ❑ *What is the timeline for achieving these objectives, and what are the critical milestones along the way?*
Success Criteria	Define Success Criteria Success criteria are the specific, measurable standards or benchmarks used to evaluate the effectiveness and impact of policies, programs, or initiatives. They define the desired outcomes and performance indicators that signify the achievement of objectives, ensuring accountability and guiding decision-making processes to enhance public service delivery.	❑ *What are the specific goals or objectives that need to be achieved?* ❑ *How will we measure progress and determine if these goals are being met?* ❑ *What are the key performance indicators (KPIs) or metrics to evaluate success?* ❑ *What are the expected outcomes or benefits for the stakeholders or community?* ❑ *What is the timeline for achieving these goals, and what milestones should be reached along the way?*

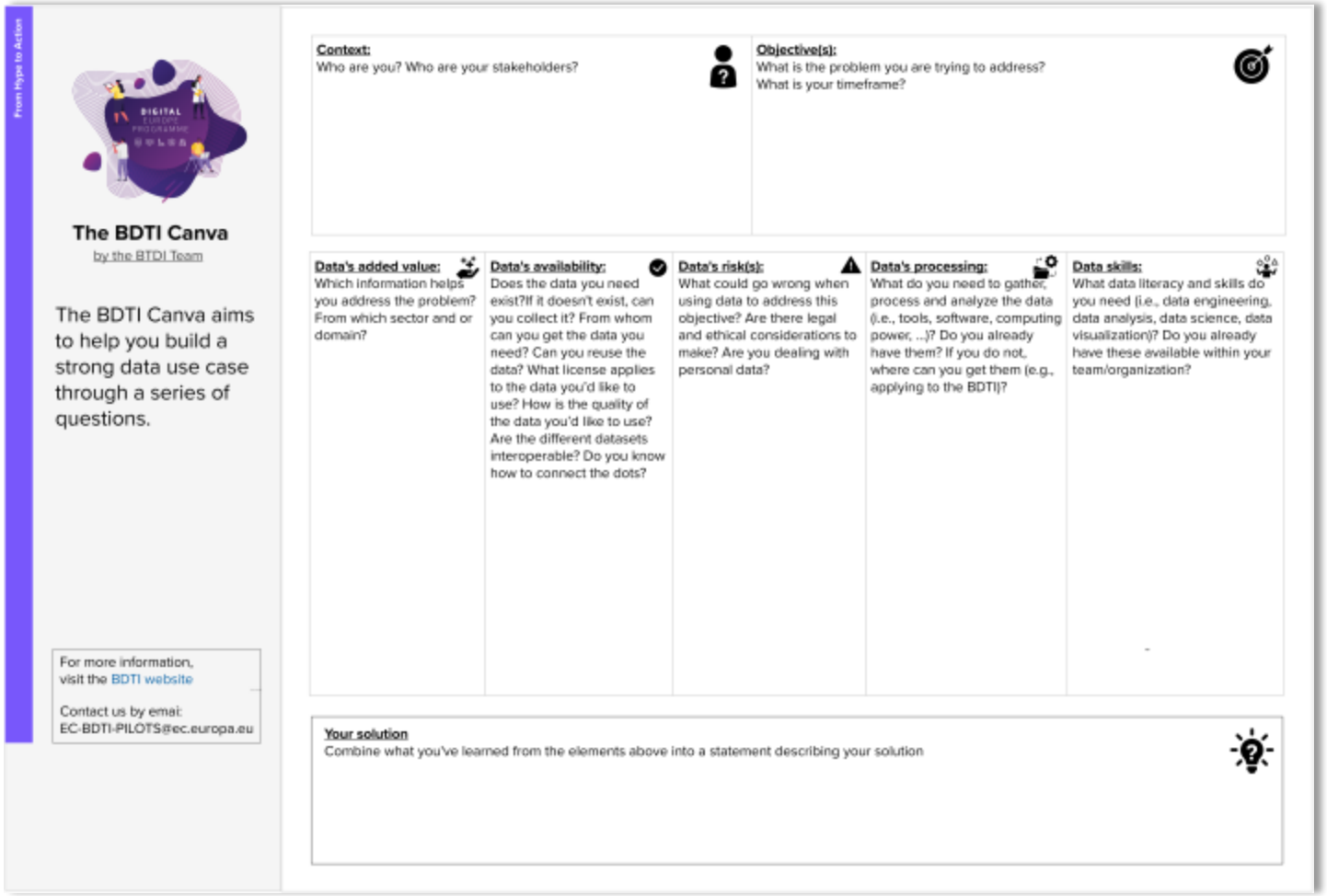
Tool: Use case readiness - Determine your next step

Clearly signifies that you should be capable of explaining this issue to a colleague in just one sentence.

- ❑ Use this tool to assess your use case readiness and determine your next step



Idea: BDTI Canva



To help you build your use case, the BDTI Team has developed the BDTI Canva to accompany to ensure that your problem has the right solution.

Discover

the BDTI Canva on the BDTI website

IDEA

You can jump to the desired page by clicking the links

Establishing the foundation

Strategic Fit

Collaboration

Skills

Does the use case fit strategic goals?

- Ensuring that public administrations **align** their data analytics initiatives **with strategic goals** fosters better resource allocation, enhances transparency, and drives more informed policy-making, ultimately leading to improved public services and greater trust from citizens.

Why collaboration?

- Leveraging **cross-departmental collaboration** enhances problem-solving efficacy.
- **Identifying stakeholders** like sponsors or data producers further supports the data analytic initiative.

What skills?

- The combination of **supporting skills** such as effective leadership combined with **technical skills** like statistical modelling, are essential foundation for public administrations to excel in data analytics.

Strategic Fit

Strategic Fit | Collaboration | Skills

Strategic Fit refers to the alignment of data projects within public administration in the European Union (EU) with the overarching strategic goals, regulatory frameworks, and environmental considerations.

This alignment ensures that data initiatives not only comply with legal requirements but also contribute to the broader objectives of transparency, innovation, economic growth, and societal benefit.

Tool: Questionnaire on Strategic Fit

How do we verify if the use case fits into a broader strategy?

Topic	*Description*	*Questions*
Policy Adherence	Policy adherence refers to the alignment between public administration's resources, capabilities, and policy objectives, laid out in relevant documents, to effectively address public needs and challenges. It ensures that public sector strategies are coherent with the internal strengths of government agencies and the external socio-political environment. Achieving strategic fit in public administration enhances the efficiency and effectiveness of public service delivery and policy implementation.	❑ *Is the project's result aligned with our entity's/EU's strategy?* ❑ *What official documents can we use to support our reasoning?* ❑ *What pillars of what EU strategy are we supporting with this project?* ❑ *Are we promoting data accessibility and reusability?* ❑ *Are we facilitating data sharing/interoperability?* ❑ *Are we adhering to the ethical practices/regulations?*
Value Identification	Value Identification involves pinpointing the specific benefits that the product delivers to its users and stakeholders. This includes identifying how the data product enhances decision-making, improves operational efficiency, and supports transparency and accountability within public sector organisations. By clearly articulating these benefits, public administrators can justify investments in data products and ensure they align with the broader goals of serving the public interest.	❑ *Has the future value of the project's product been assessed?* ❑ *Do we know what benefits it would bring? To whom? Citizens? Administration?* ❑ *What local/EU policies would the end product and its benefits support?*

Use this table to understand what strategic fit may mean for your use case

Each topic includes a non-exhaustive sample of questions to help you prepare your own list

Legal framework

Topic	*Description*
Open Data Directive (2019)	Aimed at enhancing the accessibility and reusability of public sector and publicly funded data, this directive promotes **transparency** by opening government-held data for public use. It encourages **innovation** by enabling developers and businesses to create new services using data and contributes to **economic growth** by fostering a data-driven economy.
European Data Strategy (2020)	Focused on establishing a **single market for data** across Europe, this strategy ensures the availability of data for use in society and the economy. It emphasises empowering companies and individuals by giving them **control over the data they generate**, thus balancing accessibility with privacy and ownership.
Data Governance Act (2020)	Designed to create a framework for **facilitating data sharing** across the EU, this act encourages collaboration between sectors and public entities. It aims to establish a **trustworthy and fair environment** for data exchange, reducing barriers to innovation and promoting cross-sector cooperation.
Data Act (2022)	Aims to define the rules on **who can access and use data** generated in the EU, across all economic sectors. This act clarifies the conditions under which data sharing can occur, ensuring fairness and fostering innovation while protecting the rights of those who generate the data.
Interoperable Europe Act (2024)	Seeks to create an ecosystem of **shared interoperability solutions** for the EU's public sector. By establishing a governance structure for interoperability, it facilitates collaboration and the seamless exchange of information between European public administrations. This act supports efficient digital public services across borders.
AI Act (2024)	Establishes clear **requirements and obligations** for AI developers and deployers, focusing on the ethical and safe use of AI technologies. This act provides a regulatory framework to address risks and ensure that AI deployment aligns with EU values, enhancing trust and innovation in AI applications.

Check the legal framework to be considered

IMPORTANT

Collaboration

Strategic fit | **Collaboration** | Skills

Collaboration refers to the coordinated effort of multiple stakeholders, including government agencies, private sector entities, non-governmental organisations, and citizens, to work together towards common goals.

This joint effort aims to enhance data accessibility, interoperability, and utility, while adhering to regulatory frameworks and strategic objectives. Effective collaboration ensures that data initiatives are inclusive, transparent, and capable of driving innovation and societal benefits.

Use the tables in the upcoming pages to understand what collaboration may mean for your use case

Each topic includes a non-exhaustive sample of questions to help you prepare your own list

Tool: Questionnaire on Collaboration (1/3)

How do we want to efficiently and effectively collaborate?

Topic	*Description*	*Question*
External Collaboration	External collaboration in the context of public administration involves partnerships and cooperative efforts between government agencies and external entities such as non-governmental organisations (NGOs), private sector companies, academia and community groups. This collaboration aims to leverage the unique strengths and resources of each partner to address public issues more effectively and efficiently. By working together, these entities can enhance service delivery, foster innovation, and improve outcomes for the community.	❑ *Who can the key external stakeholders involved in this collaboration be? Academia? NGOs? Startups?* ❑ *What are we getting out of this cooperation? What are they getting? Would they be inclined to collaborate at all?* ❑ *Can we join a network of cities or any other relevant network to collaborate/learn from similar cases?* ❑ *How are these stakeholders identified and engaged in the project?* ❑ *What communication channels are established for effective collaboration?* ❑ *What potential risks have been identified in this collaboration? How are these risks being mitigated and managed? Are there contingency plans in place for unforeseen challenges?* ❑ *How are ethical considerations and data privacy managed in this collaboration?* ❑ *What agreements or contracts are in place to formalise the partnership? What agreements are missing?*
Inter-departmental Collaboration	Interdepartmental collaboration refers to the cooperative efforts and coordination between different departments or divisions within an organisation. This type of collaboration aims to break down silos, enhance communication, and leverage diverse expertise to achieve common goals and improve overall organisational performance. By working together, departments can share resources, align strategies, and address complex challenges more effectively.	❑ *Which departments should be involved in this collaboration?* ❑ *Who are the key stakeholders from each department?* ❑ *How are these stakeholders identified and engaged in the project?* ❑ *What are the primary objectives of this interdepartmental collaboration?* ❑ *How do these objectives align with the overall goals of the organisation? How do they align with each department's objectives? What's in it for them? Will they be willing to help?* ❑ *What resources (financial, human, technological) will be contributed by each department? How are these resources managed and allocated throughout the project?* ❑ *What communication channels are established for effective collaboration?* ❑ *What potential risks have been identified in this collaboration?* ❑ *Are there any legal or regulatory requirements that need to be addressed (e.g. data sharing)? How are ethical considerations and data privacy managed in this collaboration?*

Tool: Questionnaire on Collaboration (2/3)

How do we want to efficiently and effectively collaborate?

Topic	*Description*	*Question*
Administration – IT Collaboration	Public administration and IT collaboration in data projects involves a collaborative decision-making process where both administration and technology engage with diverse stakeholders to make informed data and technology decisions. This partnership ensures alignment on key metrics and goals, facilitating optimal decision-making and resource allocation. By integrating civil service objectives with IT capabilities, public administration can enhance service delivery, operational efficiency, and innovation.	❑ *What processes are in place to ensure that both administration and IT are engaged with diverse stakeholders?* ❑ *How are civil service objectives integrated with IT capabilities?* ❑ *What challenges have been encountered in the collaboration, and how have they been addressed?* ❑ *Do we have the buy-in of the IT decision makers?*
Executive Commitment	Leaders understand importance of data's impact on the future of the public administration and are active in investing in and holding the organisation accountable for change. They provide strategic direction, resources, and governance to ensure the project's alignment with broader goals. They actively engage stakeholders, monitor performance, and advocate for the product's value to ensure sustained support and adoption. Additionally, they manage risks and enforce policies to maintain data privacy, security, and compliance. The leadership's support at various levels is essential for a project to succeed	❑ *Who are my key stakeholders? Data officers, department heads, city council, mayor/president, etc.?* ❑ *Do the key stakeholders support this initiative?* ❑ *Are they actively participating in promoting it?* ❑ *If not, should they? How do we onboard them?*

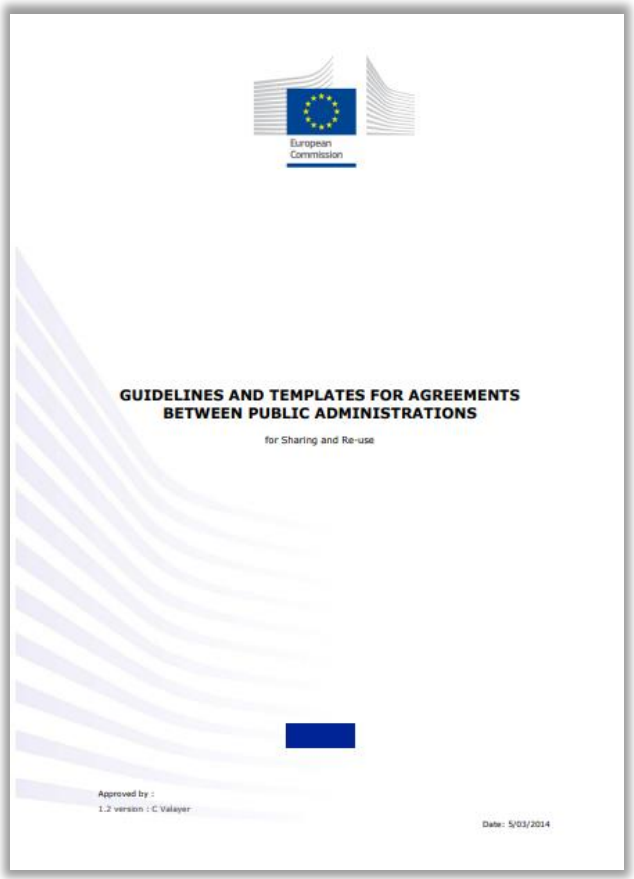
Tool: Questionnaire on Collaboration (3/3)

How do we want to efficiently and effectively collaborate?

Topic	**Description**	*Question*
Delivery Commitment	A delivery commitment refers to the formal pledge by a government entity to allocate necessary resources, including personnel, budget, and technology, to develop and implement a data solution. This commitment ensures that the project has the required support and oversight to meet its objectives and deliverables within the specified timeframe. It also involves accountability measures to track progress and ensure that the data solution effectively addresses the public needs it was designed to solve.	❑ *Do we have the commitment of particular departments (e.g. IT, data, analytics, etc.) to provide help?* ❑ *Did these departments agree to provide resources to complete the project?* ❑ *Even if they did, did we get the buy-in of the stakeholders who could influence the withdrawal of the agreement?*
Approval Processes	The approval process for conceiving and working on a data product in an entity. Takes into consideration the time needed for getting all the required approvals for starting a project but also making important decisions during its delivery. Includes the need for superiors' approvals, issuing public tenders and review by other public/government institutions whenever needed.	❑ *What are the steps involved in the approval process for starting a data product project?* ❑ *How much time is typically required to obtain all necessary approvals?* ❑ *What are the key milestones in the approval process?* ❑ *Are there specific criteria that must be met for approval at each stage?* ❑ *How are public tenders issued and managed as part of the approval process?* ❑ *What role do other public or governmental institutions play in the approval process?* ❑ *How are important decisions made and approved during the project's delivery? What documentation is required at each stage of the approval process?*

Check how to identify and manage stakeholders on the next page

Framework for public administration collaboration



The document „Guidelines and templates for agreements between public administrations for Sharing and Re-use" is part of the ISA Action **"Sharing and Re-use Strategy"** which aims to create a holistic approach for **sharing and re-using public service delivery solutions** across Europe. The strategy seeks to optimise sharing and re-use activities, increasing efficiency and savings for public administrations.

The document provides examples of **templates** and **guidelines** for agreements, including:

- a comprehensive **collaboration framework**,
- guidelines for **shared services agreements**,
- and templates for developing and **re-using shared tools**.

IMPORTANT

Idea: Inspiration for intergovernmental agreements

Helsinki Region Infoshare (HRI)

This initiative involves municipalities in the Helsinki region (Helsinki, Espoo, Vantaa, and Kauniainen) sharing open data to improve transparency and innovation. The data include information on transportation, environment, and public services, which is made available through a common portal.

Ghent and Bruges Local Open Data Economy

LODE is a project by the cities of Ghent and Bruges that explores the social and economic potential of Open Data. The project aims to understand how Open Data can make a significant impact in the next decade. The study identifies challenges such as the lack of visibility on data needs, a non-committal attitude towards Open Data, and the need for a clear vision and investments in Open Data operations.

Smart Dublin Initiative

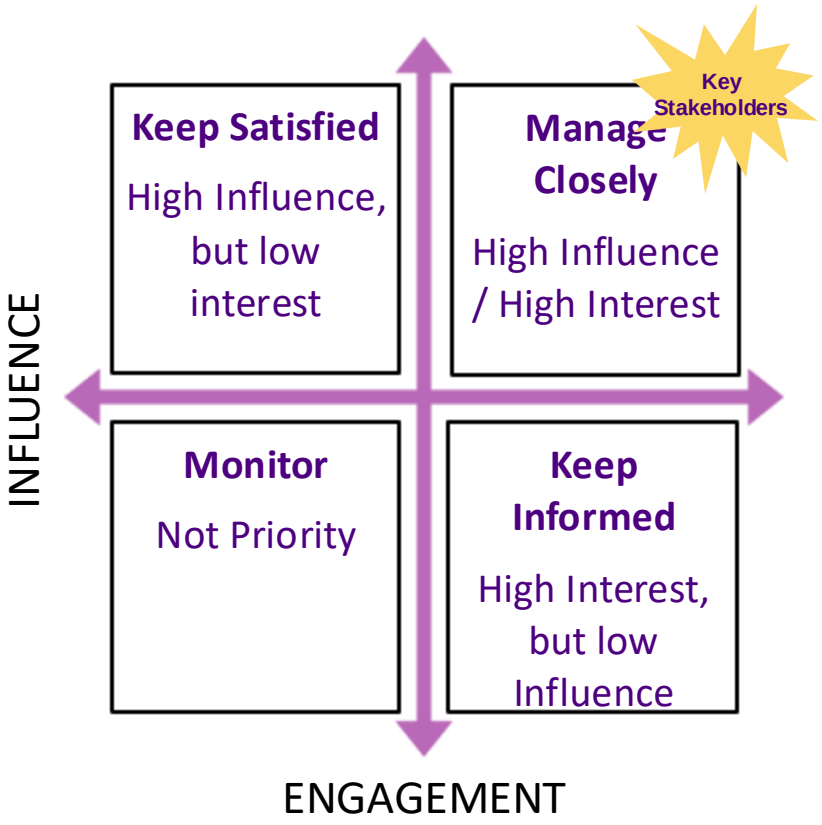
This collaboration involves the four Dublin local authorities (Dublin City Council, Fingal County Council, South Dublin County Council, and Dún Laoghaire-Rathdown County Council) sharing data to address urban challenges. The initiative focuses on areas such as traffic management, waste collection, and public safety, using data to develop innovative solutions.

IDEA

Tool: Stakeholder mapping

How can we map our stakeholders to target our collaboration and communication efforts?

The Influence/Engagement grid is a tool that maps stakeholders, to help you identify the most important people to engage with, and to define how to collaborate with each.



Stakeholder Mapping Tips and Tricks:

To effectively collaborate with your stakeholders, you can use the **Engagement vs. Influence Stakeholder Mapping Grid**. This tool helps you to categorise stakeholders based on their level of engagement and influence, allowing you to plan your collaboration based on the four quadrants.

X-Axis (Engagement): Measures the level of interest or engagement a stakeholder has in the project.
Y-Axis (Influence): Measures the level of influence or power a stakeholder has over the project.

Here are some examples of how you could define your interaction with the stakeholders per quadrant:

Keep Satisfied
- Engage and consult

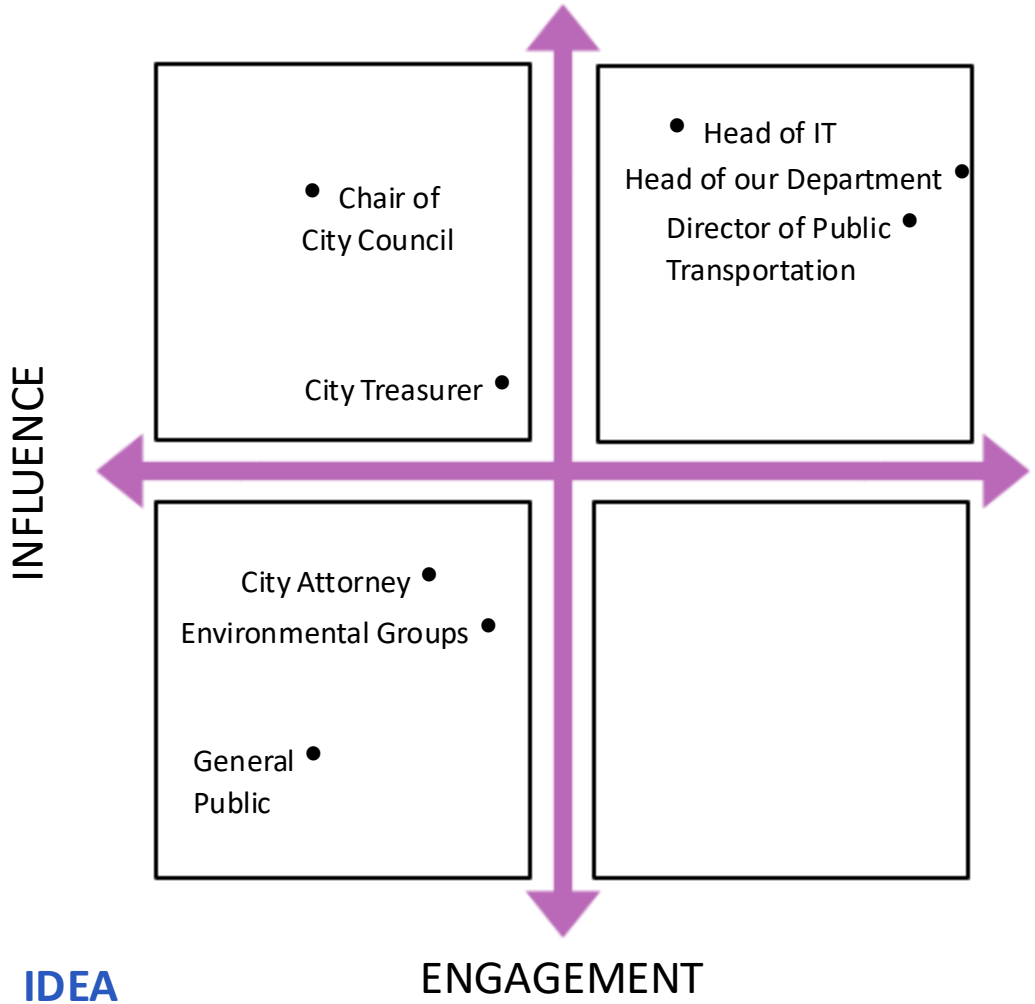
Try to increase engagement and move into the box to the right

Manage Closely
- Include in Governance and decision-making
- Engage directly and consult regularly

Monitor
- Inform with basic communication: newsletters, emails, blogs etc.
- Enhance engagement by moving into the right-hand box

Keep informed
- Use interest to get involved and consult in areas of expertise
- Use as supporters and ambassadors

Idea: Stakeholder mapping for a Public Transportation Efficiency Dashboard project



The dashboard will provide real-time data and analytics on public transportation usage, efficiency, and areas for improvement.

Stakeholder	*Interest*	*Engagement Strategy*
Head of our Department	*Very High* - removing obstacles, ensuring budget	Progress reports, weekly meetings
Head of IT Department	*High* - Infrastructure and technical support	Bi-weekly technical meetings, support tickets
Director of Public Transportation	*High* - Provides operational data and insights	Workshops, data sharing agreements
Chair of City Council	*Medium* - Approval and oversight	Monthly presentations, progress reports
City Treasurer	*Medium* - Budget allocation and financial oversight	Quarterly budget reviews, financial reports
Environmental Groups	*Medium* - Data on emissions and sustainability	Periodic updates, collaborative meetings
City Attorney	*Medium* - Compliance and data privacy	Compliance reviews, legal consultations
General Public	*Low* - Occasional interest in project updates	Public announcements, community meetings

Assessment of the team's technical skills

Strategic fit | Collaboration | **Skills**

To develop a data product on BDTI, a subset of the presented skills is required. Note that the actual team setup will vary from case to case. This section will help you decide what roles and skills you may need for your use case.

Technical Skills

- Key technical skills in this domain include a strong foundation in **statistical modelling** and **machine learning**, which are essential for analysing complex datasets and deriving actionable insights.
  Roles: Data Scientist
- Proficiency in **data management** is critical and covers skills such as data collection, refinement, and cleaning, along with a robust understanding of database systems like SQL and No-SQL.
  Roles: Data Engineer, Data Architect
- **Programming skills**, particularly in scripting languages such as Python, are crucial for automating data analysis tasks and developing custom analytical tools.
  Roles: Backend Developer
- Equally important are **frontend developing** and **visualisation skills**. The ability to translate data-driven insights into clear and compelling narratives, supported by interactive dashboards and infographics.
  Roles: Frontend Developer, Data Visualisation Specialist
- For more advanced workflows, **automating and streamlining the software development** and deployment processes may be necessary.
  Roles: DevOps, MLOps, Cloud Architect

Are there any gaps in terms of skills that would require outsourcing?

Source: Data Playbook, 'Understanding how data matters', p. 69-70, https://preparecenter.org/toolkit/data-playbook-toolkit-v1/

Tool: Technical skills assessment – Data Management

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
Data Engineers	Data Engineers design, build, and maintain the infrastructure and systems that allow for the collection, storage, and analysis of large volumes of data. They ensure data are accessible, reliable, and efficiently processed for use by data scientists and analysts.	❑ *SQL and NoSQL Databases: Proficiency in database management and querying.* ❑ *ETL (Extract, Transform, Load) Processes: Experience with data integration and transformation.* ❑ *Big Data Technologies: Knowledge of Hadoop, Spark, and other big data frameworks.* ❑ *Programming Languages: Proficiency in Python, Java, or Scala.* ❑ *Data Warehousing Solutions: Experience with tools like Amazon Redshift, Google BigQuery, or Snowflake.*
Data Architects	Data Architects design and manage the data infrastructure of an organisation. They create blueprints for data management systems, ensuring data are organised, stored, and accessed efficiently and securely.	❑ *Data Modelling: Proficiency in designing data models and schemas* ❑ *Database Management: Experience with SQL and NoSQL databases.* ❑ *Big Data Technologies: Knowledge of Hadoop, Spark, and other big data frameworks.* ❑ *Cloud Platforms: Experience with AWS, Azure, or Google Cloud.* ❑ *ETL Processes: Skills in data integration and transformation.*

Use this table to understand what DATA MANAGEMENT may mean for your use case

Tool: Technical skills assessment – Frontend/Visualisation

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
Data Visualisation Specialist	Data Visualisation Specialists create visual representations of data to help stakeholders understand complex information. They use tools and techniques to design charts, graphs, and dashboards that make data insights accessible and actionable.	❑ *Data Visualisation Tools: Proficiency in Tableau, Power BI, or D3.js.* ❑ *Design Principles: Understanding of visual design and user experience.* ❑ *Data Analysis: Ability to analyse and interpret data.* ❑ *Programming Languages: Skills in Python or R for data manipulation.* ❑ *Communication: Strong ability to present data insights effectively*
Frontend Developers	Frontend Developers build the user interface and user experience of web applications. They use technologies like HTML, CSS, and JavaScript to create responsive and interactive websites that provide a seamless experience for users.	❑ *HTML/CSS: Proficiency in web markup and styling.* ❑ *JavaScript: Strong skills in JavaScript and frameworks like React, Angular, or Vue.js.* ❑ *Responsive Design: Experience with creating responsive and mobile-friendly designs.* ❑ *Version Control: Proficiency in Git and GitHub.* ❑ *Cross-Browser Compatibility: Skills in ensuring compatibility across different browsers.*

Use this table to understand what FRONTEND AND VISUALISATION may mean for your use case

Tool: Technical skills assessment – Math and Statistics

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
Data Scientists	Data Scientists analyse and interpret complex data to help organisations make informed decisions. They use statistical methods, machine learning, and data visualisation techniques to uncover patterns, trends, and insights from data.	❑ *Statistical Analysis: Strong foundation in statistics and probability.* ❑ *Machine Learning: Proficiency in algorithms, model building, and evaluation.* ❑ *Programming Languages: Proficiency in Python or R.* ❑ *Data Visualisation: Skills in tools like Matplotlib, Seaborn, or Tableau.* ❑ *Data Wrangling: Ability to clean and preprocess data effectively.*

Use this table to understand what MATH AND STATISTICS may mean for your use case

Tool: Technical skills assessment – Programming

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
Backend Developers	Backend Developers focus on the server-side logic, databases, and APIs that power web applications. They ensure that the backend systems are robust, scalable, and secure, enabling the frontend to function smoothly.	❑ *Server-Side Languages: Proficiency in languages like Java, Python, Ruby, or Node.js.* ❑ *Database Management: Experience with SQL and NoSQL databases.* ❑ *API Development: Skills in creating and managing RESTful or GraphQL APIs.* ❑ *Security Best Practices: Knowledge of securing backend services.* ❑ *Version Control: Proficiency in Git and GitHub.*

Use this table to understand what PROGRAMMING may mean for your use case

Tool: Technical skills assessment – Operations

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
DevOps	DevOps professionals bridge the gap between development and operations teams by automating and streamlining the software development and deployment process. They focus on improving collaboration, continuous integration, and delivery to enhance the efficiency and reliability of software systems.	❑ *CI/CD Pipelines: Experience with continuous integration and continuous deployment tools.* ❑ *Infrastructure as Code (IaC): Proficiency in tools like Terraform or Ansible.* ❑ *Containerisation: Knowledge of Docker and Kubernetes.* ❑ *Monitoring and Logging: Experience with tools like Prometheus, Grafana, or ELK stack.* ❑ *Scripting Languages: Proficiency in Bash, Python, or PowerShell.*
MLOps	MLOps (Machine Learning Operations) specialists integrate machine learning models into production environments, ensuring they are scalable, reliable, and maintainable. They work on automating the deployment, monitoring, and management of machine learning models.	❑ *Model Deployment: Experience with deploying machine learning models in production.* ❑ *CI/CD for ML: Knowledge of continuous integration and deployment for machine learning.* ❑ *Monitoring and Maintenance: Skills in monitoring model performance and retraining.* ❑ *Data Versioning: Proficiency in tools like DVC or MLflow.* ❑ *Cloud Platforms: Experience with AWS, Azure, or Google Cloud for ML operations.*
Cloud Architects	Cloud Architects design and oversee the cloud computing strategy of an organisation. They are responsible for planning, implementing, and managing cloud infrastructure and services to ensure scalability, security, and cost-efficiency.	❑ *Cloud Platforms: Proficiency in AWS, Azure, or Google Cloud.* ❑ *Infrastructure as Code (IaC): Experience with Terraform, CloudFormation, or ARM templates.* ❑ *Security Best Practices: Knowledge of cloud security and compliance.* ❑ *Networking: Skills in cloud networking and connectivity.* ❑ *Cost Management: Ability to manage and optimise cloud costs.*

Use this table to understand what OPERATIONS may mean for your use case

Idea: Example Use Cases with required Technical Roles

Example 1

Use Case:
Display aggregated city traffic data on a map from a pre-prepared csv file with all the necessary information in a structured form.

Technical roles required:
- Data visualisation specialist – creates maps/dashboards to display data.
- Junior data engineer – creates data storage and loads data.

Example 2

Use Case:
Build a data platform that would collect air pollution data in various formats from external sources and make them available for download or display in dashboards in real time.

Technical roles required:
- Data architect – prepares the data model accommodating various sources.
- Data engineer – implements the data pipelines that allow the ingestion of the data to the platform, transformations and storage for the final consumption.
- Backend developer – implements the business logic, helps the data engineer in creating the data transformations and consumption solutions.
- Data visualisation specialist/frontend developer – creates UIs/dashboards for data upload and consumption.
- DevOps – automates the integration, testing and deployment of the solution.

These examples show that per use case, only a part of the listed technical roles are required

IDEA

Assessment of the team's supporting skills

To effectively and efficiently deliver a project, leadership and management must complement the technical skills. Note that the actual team setup will vary from case to case. This section will help you decide what skills you may need for your use case.

Supporting Skills:

- **Effective leadership** is crucial, as it ensures that teams are guided with a clear vision and purpose.
- **People management** and human resources expertise are fundamental to building and sustaining a competent and motivated team.
- **Finance and accounting** knowledge is essential for budgeting, financial planning, and ensuring that the project is economically viable.
- **Domain knowledge** is vital for an in-depth understanding of public administration challenges and regulations, enabling effective decision-making and strategy development.
- Clear, concise, and transparent **communication** ensures that ideas, goals, and expectations are properly conveyed across the team and stakeholders.
- **Project management** expertise is vital for planning, executing, and closing projects effectively, ensuring that objectives are met on time and within budget.
- **Policy expertise** allows to develop, analyse, and implement data products that align with governmental objectives and public interests, ensuring legal compliance and social equity.

Are there any gaps in terms of skills that would require outsourcing?

Source: Data Playbook, 'Understanding how data matters', p. 69-70, https://preparecenter.org/toolkit/data-playbook-toolkit-v1/

Tool: Supporting skills assessment (1/2)

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
Project Manager	A project manager in the public administration context is responsible for planning, executing, and overseeing projects that aim to improve public services and infrastructure. They ensure that projects are completed on time, within budget, and meet the required standards and regulations. This role involves coordinating with various stakeholders, managing resources, and mitigating risks to achieve the desired outcomes.	❑ *Strong organisational and planning abilities* ❑ *Excellent communication and interpersonal skills* ❑ *Proficiency in project management software and tools* ❑ *Ability to manage budgets and financial resources* ❑ *Problem-solving and critical thinking skills*
Delivery Lead	A delivery lead in the public administration context is responsible for ensuring the successful delivery of projects and services that meet the needs of the community. They oversee the implementation process, manage teams, and ensure that deliverables are completed on time and within scope. This role involves close collaboration with stakeholders, continuous monitoring of progress, and addressing any issues that may arise to ensure smooth project execution.	❑ *Strong leadership and team management skills* ❑ *Excellent communication and stakeholder engagement abilities* ❑ *Proficiency in project and service delivery methodologies* ❑ *Ability to manage budgets and ensure timely delivery* ❑ *Problem-solving and adaptability skills*
Business Analyst	A business analyst in the public administration context is responsible for analysing and improving processes, systems, and policies to enhance public services and operations. They gather and interpret data, identify areas for improvement, and provide recommendations to optimise efficiency and effectiveness. This role involves collaborating with various departments, understanding stakeholder needs, and ensuring that proposed solutions align with organisational goals.	❑ Strong analytical and critical thinking abilities ❑ Excellent communication and interpersonal skills ❑ Proficiency in data analysis and visualisation tools ❑ Ability to understand and document business processes ❑ Problem-solving and decision-making skills

Use this table to understand what SUPPORTING SKILLS your use case may need

Tool: Supporting skills assessment (2/2)

Do we need those skills? Do we have them in the team?

Role	*Description*	*Skills Checklist*
Subject Matter Expert	A subject matter expert (SME) is an individual with extensive knowledge and expertise in specific areas of government operations, policies, and procedures. They provide valuable insights and guidance on best practices, regulatory compliance, and effective governance strategies. SMEs play a crucial role in advising public officials, developing policy frameworks, and ensuring the efficient and transparent functioning of public institutions.	❑ *Policy Analysis and Development* ❑ *Domain-Specific Knowledge* ❑ *Analytical and Critical Thinking* ❑ *Communication Skills* ❑ *Ethical Judgment and Integrity*

Use this table to understand what SUPPORTING SKILLS your use case may need

Idea: BDTI Skills Studio

The BDTI Skills Studio is a learning space designed for public administration professionals to develop essential data skills, uncover insights, and create a better future for citizens on a local, regional, and national level.

Discover

The BDTI Skills Studio

IDEA

You can jump to the desired page by clicking the links

Assessing data quality

Data assessment is a critical step in the data product creation process for public administration. It involves evaluating, inter alia, the quality, accuracy, and completeness of data used to develop the final product.

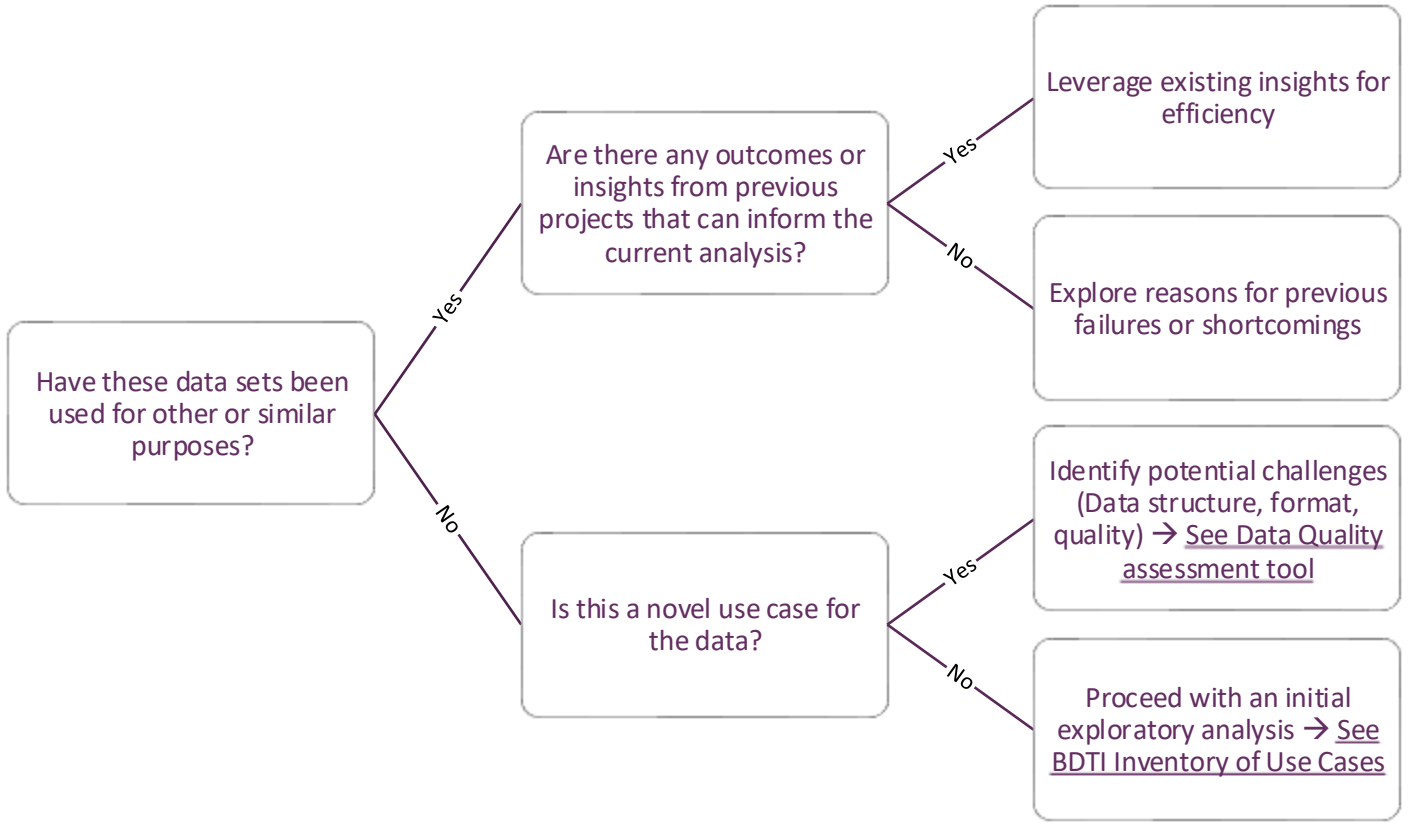
By **identifying inconsistencies, errors, or gaps**, data assessment ensures that the **end-product meets the required standards** for effective policy-making and public service delivery, thereby supporting **transparent and efficient governance**.

Source: Data Playbook, 'Making Data Useful, Useable and Shareable', p. 303, https://preparecenter.org/toolkit/data-playbook-toolkit-v1/

Pre-Assessment: Have these data sets been used before?

Understanding the previous usage and outcomes of data sets can significantly enhance the efficiency and effectiveness of your data analysis.

- ❑ Use this tool to ensure your work is based on a more informed and strategic approach
- ❑ Explore the European Data Portal, where almost 2M datasets are available for download through inter alia SPARQL or REST API



IMPORTANT

Tool: Operational data assessment

What operational data aspects should be considered?

Role	*Description*	*Question*
Continuity	Continuity in data refers to the seamless and uninterrupted flow of data over time, ensuring that data are consistently available and up-to-date for analysis and decision-making. Refers also to the fact of data being available over time (i.e. the data source being continuously maintained).	❑ *What is the certainty data will be available going forward?* ❑ *How reliable is the data source in terms of SLAs (up-time, history, format backward compatibility, etc.)?* ❑ *Are there any known issues with data latency or timeliness?* ❑ *Is financial commitment required now/is it foreseen in the future?* ❑ *What risks are associated with data discontinuity, and how are they mitigated?*
Sensitivity	Data sensitivity refers to the classification of data based on its level of importance and the potential impact of its exposure. Sensitive data typically includes information that, if disclosed, could cause harm to individuals, organisations, or systems. This can include personal information, financial records, intellectual property, and other confidential or proprietary information.	❑ *Are there any compliance requirements (e.g., GDPR) that need to be considered?* ❑ *Can the data be shared/displayed in the application?* ❑ *Do we need to introduce masking/aggregation mechanisms?* ❑ *How will we make sure we don`t disclose inappropriate data, i.e. what technology and processes will we use?* ❑ *What are the data retention policies for your organisation?* ❑ *How long will the data need to be stored, and what are the implications for storage capacity?*
Quality	Data quality refers to the condition of a set of values of qualitative or quantitative variables that are accurate, complete, reliable, and relevant to the intended use. High data quality ensures that the data are fit for its intended purposes in operations, decision-making, and planning. Poor data quality can lead to incorrect conclusions, inefficiencies, and potential risks in business processes.	❑ *What is the data quality at the source? How common are errors? Are the errors significant?* ❑ *What validation processes are needed to ensure data conform to the expected format and structure?* ❑ *How much effort would it be to remove errors from incoming data?*

Use this table to assess the operational aspect of data for your use case

For a detailed data sensitivity and quality assessment, check out the next pages

Tool: Technical data assessment

What technical data aspects should be considered?

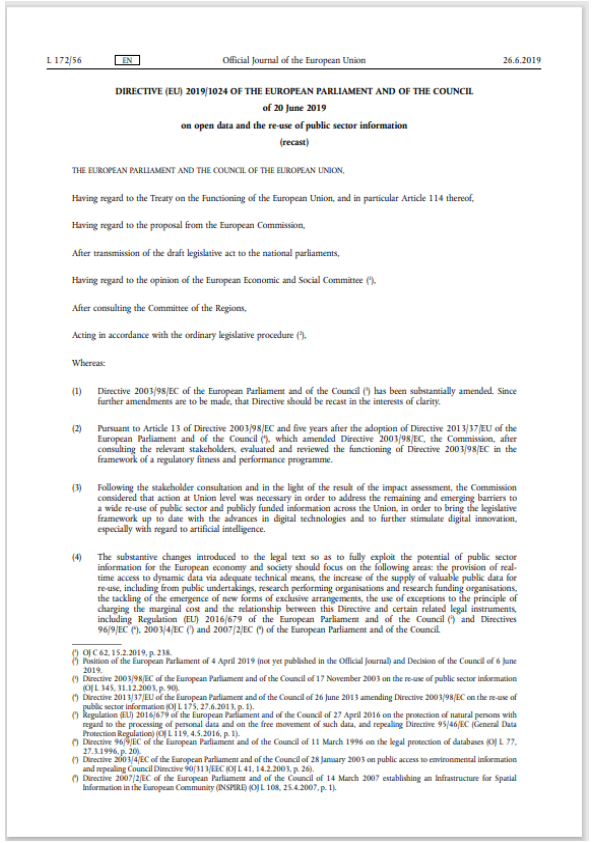
Role	*Description*	*Questions*
Delivery	Delivery pertains to the methods and processes used to transport data from one location to another, ensuring that it reaches its intended destination accurately and efficiently. The most common ways are automated batches (a bulk upload of the data), automated streaming (uploading data one by one, as they become available) or manual upload (a batch upload of a selected chunk of data).	❑ *What is the delivery method of the data? Batch? Streaming? Manual upload?* ❑ *How frequently will the data need to be delivered (e.g., real-time, hourly, daily)?* ❑ *How easy is it to integrate the data delivery method with the BDTI platform?* ❑ *How many distinct sources will there be? What would be the workload to integrate all data sources?*
Volume	Volume describes the amount of data being generated, stored, and processed, often measured in terms of bytes (and its multipliers), and is a key aspect of big data.	❑ *What is the expected volume of data to be transferred?* ❑ *How will the volume of data impact processing times and performance?* ❑ *What are the cost implications of storing and processing large volumes of data?* ❑ *How will you manage and optimise costs related to data volume?*
Format	Format refers to the structure and organisation of data, which can vary widely (e.g., CSV, JSON, XML, Avro, etc.) and affects how data are stored, processed, and interpreted.	❑ *What is the format of the data? Are there various formats?* ❑ *Does BDTI natively support storing this format?* ❑ *What (and how costly) transformations would need to be done on these data to store it?* ❑ *Does BDTI natively support processing this format?* ❑ *What (and how costly) transformations would need to be done on these data to use/display it?*

Use this table to assess the technical aspect of data for your use case

For a detailed data sensitivity and quality assessment, check out the next pages

Legal: Access and reuse of open data

Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information



L 172/56 EN Official Journal of the European Union 26.6.2019

DIRECTIVE (EU) 2019/1024 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 20 June 2019
on open data and the re-use of public sector information
(recast)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 114 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee (1),

After consulting the Committee of the Regions,

Acting in accordance with the ordinary legislative procedure (2),

Whereas:

(1) Directive 2003/98/EC of the European Parliament and of the Council (3) has been substantially amended. Since further amendments are to be made, that Directive should be recast in the interests of clarity.

(2) Pursuant to Article 13 of Directive 2003/98/EC and five years after the adoption of Directive 2013/37/EU of the European Parliament and of the Council (4), which amended Directive 2003/98/EC, the Commission, after consulting the relevant stakeholders, evaluated and reviewed the functioning of Directive 2003/98/EC in the framework of a regulatory fitness and performance programme.

(3) Following the stakeholder consultation and in the light of the result of the impact assessment, the Commission considered that action at Union level was necessary in order to address the remaining and emerging barriers to a wide re-use of public sector and publicly funded information across the Union, in order to bring the legislative framework up to date with the advances in digital technologies and to further stimulate digital innovation, especially with regard to artificial intelligence.

(4) The substantive changes introduced to the legal text so as to fully exploit the potential of public sector information for the European economy and society should focus on the following areas: the provision of real-time access to dynamic data via adequate technical means, the increase of the supply of valuable public data for re-use, including from public undertakings, research performing organisations and research funding organisations, the tackling of the emergence of new forms of exclusive arrangements, the use of exceptions to the principle of charging the marginal cost and the relationship between this Directive and certain related legal instruments, including Regulation (EU) 2016/679 of the European Parliament and of the Council (5) and Directives 96/9/EC (6), 2003/4/EC (7) and 2007/2/EC (8) of the European Parliament and of the Council.

(1) OJ C 62, 15.2.2019, p. 238.
(2) Position of the European Parliament of 4 April 2019 (not yet published in the Official Journal) and Decision of the Council of 6 June 2019.
(3) Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information (OJ L 345, 31.12.2003, p. 90).
(4) Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information (OJ L 175, 27.6.2013, p. 1).
(5) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016, p. 1).
(6) Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (OJ L 77, 27.3.1996, p. 20).
(7) Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC (OJ L 41, 14.2.2003, p. 26).
(8) Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) (OJ L 108, 25.4.2007, p. 1).

The directive is based on the general principle that public and publicly funded data should be reusable for commercial or non-commercial purposes.

Summary

Full Text

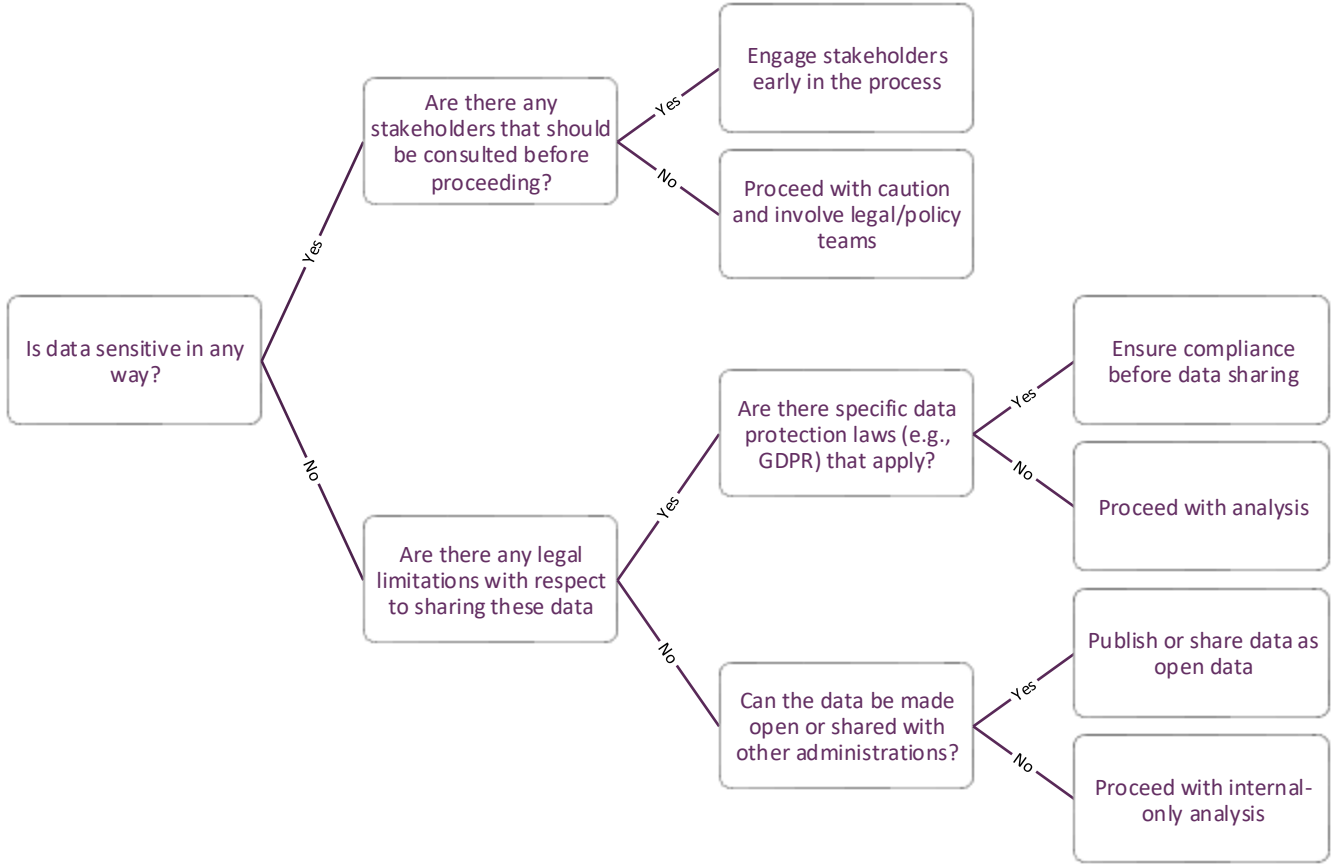
IMPORTANT

Tool: Data Sensitivity assessment

It's important to consider data sensitivity when reusing data sets. Legal limitations, such as data protection laws, must be reviewed to ensure compliance.

Depending on these factors, data can be shared openly or used internally (with periodic legal reviews).

- ❑ Use this tool to ensure data sharing is both compliant and strategically sound
- ❑ If unsure, check the EU data protection rules

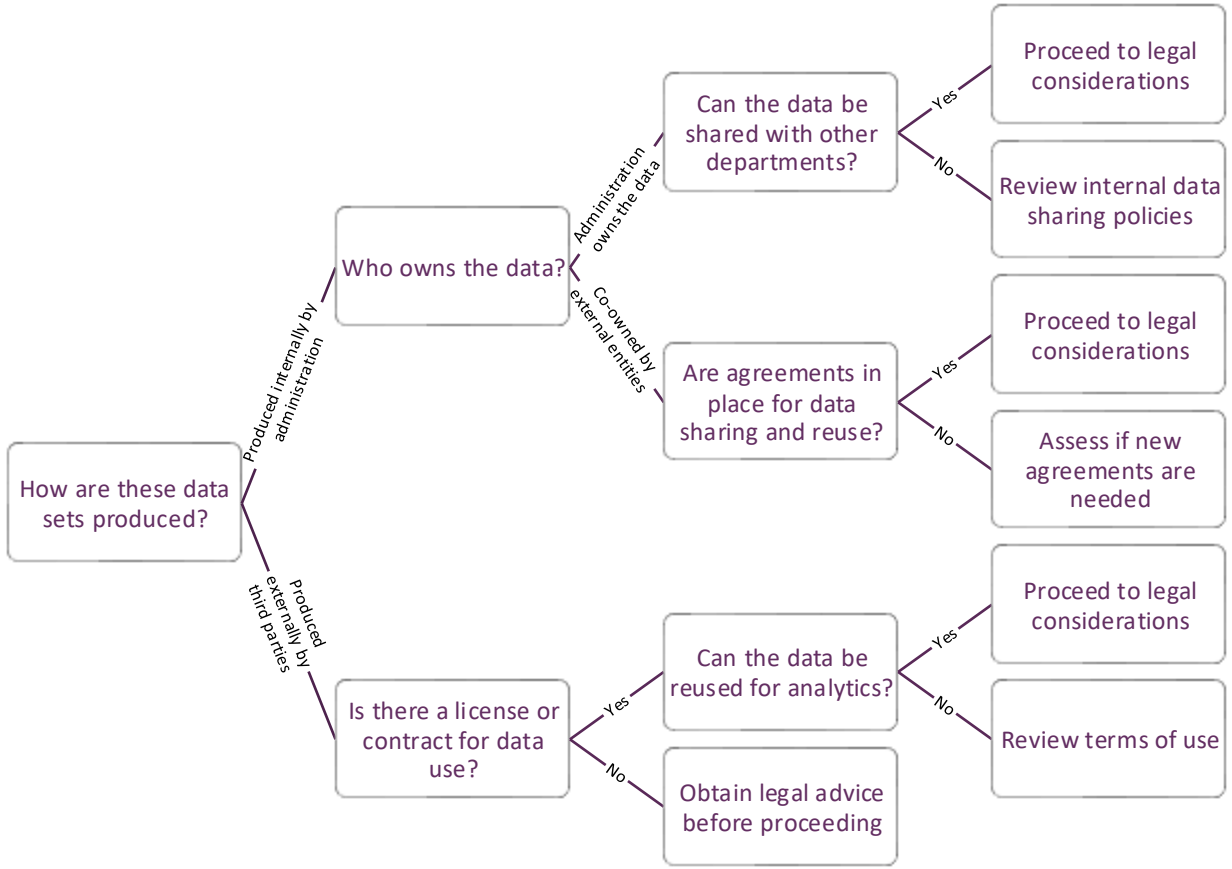


Tool: Data Compliance assessment

Assessing how data sets are produced is crucial for public administrations because it helps ensure data integrity, ownership, and compliance with legal standards.

By understanding the origin of the data, public administrations can better manage data-sharing agreements, protect sensitive information, and ensure that the data can be used effectively for data analytics and decision-making.

- ❑ Use this tool to evaluate the ownership of the data sets



Data quality dimensions for INPUT data

Data quality dimensions are criteria used to **assess the quality of input data**, ensuring they are suitable for their intended use. These dimensions help identify and measure various aspects of data quality, such as its accuracy, completeness, and consistency, to ensure that data are **reliable and valuable for decision-making processes**. By evaluating data against these dimensions, organisations can maintain **high standards of data integrity and usability**.

Accuracy

The degree to which data correctly describe the real-world object or event it represents

Completeness

The extent to which all required data are present

Consistency

The degree to which data are uniform and free from contradictions across different datasets

Timeliness

The extent to which data are up-to-date and available within the required time frame

Validity

The degree to which data conform to the defined formats, rules, and constraints

Uniqueness

Ensuring that each record is unique and not duplicated within the dataset

Integrity

The degree to which data are coherent and maintain relationships between elements

Tool: Assessment on Accuracy

Accuracy of datasets is vital for public administrations as it ensures reliable decision-making and effective use of data.

Regular validation, audits, and error rate assessments help maintain data quality and identify inaccuracies, enabling better analytics and informed policy implementation.

- ❑ Use this tool to assess the accuracy of datasets and identifying error rates to maintain data quality.

Correct Representation

Do the data correctly describe the real-world object or event they represent?

- Yes → There is correct data representation.
- No → Verify the data against reliable sources or real-world observations. Conduct regular audits and validation checks to ensure accuracy.

Example

- A dataset of the European Commission's employees accurately lists "John Smith" as an employee in the "Finance Department." — Correct
- The same dataset lists "John Smith" as an employee in the "IT Department" when he actually works in the Finance Department. — Incorrect

Error Rate

Is the error rate within acceptable limits?

- Yes → Error rate is within acceptable limits.
- No → Identify and correct errors by implementing data validation rules and automated error detection mechanisms. Regularly review and update these rules.

Example

- A dataset of public contracts has an error rate of less than 1%. — Correct
- A dataset of public contracts has an error rate of 10%, with many incorrect contract amounts and dates. — Incorrect

Tool: Assessment on Completeness

Ensuring data presence and coverage is essential for effective dataset use. Address missing elements, such as compliance data in regulatory datasets, through collection or imputation.

Verify datasets cover key areas. Regular reviews ensure completeness and alignment with evolving needs.

- ❑ Use this tool to identify missing data elements, ensure mandatory fields are populated, and verify that datasets comprehensively cover all relevant aspects of the subject matter.

Required Data Presence

Are all required data present in the dataset?

- Yes → All required data are present in the dataset.
- No → Identify missing data elements and fill in the gaps by collecting additional data or using imputation techniques. Ensure that all mandatory fields are populated.

Example

- A dataset of public projects includes all required fields such as project name, start date, end date, budget, and project manager. — Correct
- The same dataset only includes information on healthcare and education, omitting transportation and social services. — Incorrect

Data Coverage

Does the dataset cover all necessary aspects of the subject matter?

- Yes → The dataset covers all necessary aspects.
- No → Review the dataset against the requirements and ensure comprehensive data collection. Update the data collection process to capture all relevant information.

Example

- A dataset of public services includes information on healthcare, education, transportation, and social services. — Correct
- The same dataset only includes information on healthcare and education, omitting transportation and social services. — Incorrect

Tool: Assessment on Consistency

Ensure data consistency by verifying uniformity across datasets and resolving contradictions through manual or automated checks.

Standardise formats and definitions, correcting non-compliant records to maintain reliability and compatibility.

- ❑ Use this tool to ensure data uniformity by reconciling inconsistencies and enforcing standard formats and definitions across datasets for improved consistency.

Uniformity Across Datasets

Are the data uniform and free from contradictions across different datasets?

- Yes → The dataset is uniform and free from contradictions across different datasets.
- No → Manually check for data consistency across databases or implement an automated check.

Example

- Correct: Employee names and IDs are consistent across HR and payroll datasets.
- Incorrect: An employee is listed as "Jane Doe" in the HR dataset and "J. Doe" in the payroll dataset.

Standardisation

Are data formats and definitions standardised across datasets?

- Yes → Data formats and definitions are standardised across datasets.
- No → Establish and enforce data standards and definitions for the dataset. Check and modify records that do not follow the rules set.

Example

- Correct: Dates are consistently formatted as "YYYY-MM-DD" across all datasets.
- Incorrect: Dates are formatted as "DD/MM/YYYY" in one dataset and "MM-DD-YYYY" in another.

Tool: Assessment on Timeliness

Currency of data refers to the extent to which information is current, accurate, and updated within the required timeframe to meet its intended purpose effectively.

Data availability ensures that information is consistently accessible and retrievable whenever needed, supported by reliable systems or mechanisms to prevent delays or disruptions.

- ❑ Use this tool to ensure data are up-to-date with automated refresh schedules and real-time updates, and to optimise accessibility through efficient storage and retrieval processes.

Data Currency

Are the data up-to-date and available within the required time frame?

- Yes → The data are up-to-date and available within the required timeframe.
- No → Identify the root cause and implement a real-time update mechanisms or contact the responsible entity to have data updated regularly.

Example

- A dataset of public health statistics is updated monthly. — Correct
- The same dataset has not been updated in over a year. — Incorrect

Data Availability

Are the data available when needed?

- Yes → The data are available when needed.
- No → Ensure data availability by either using an API or creating a place to upload data by data providers.

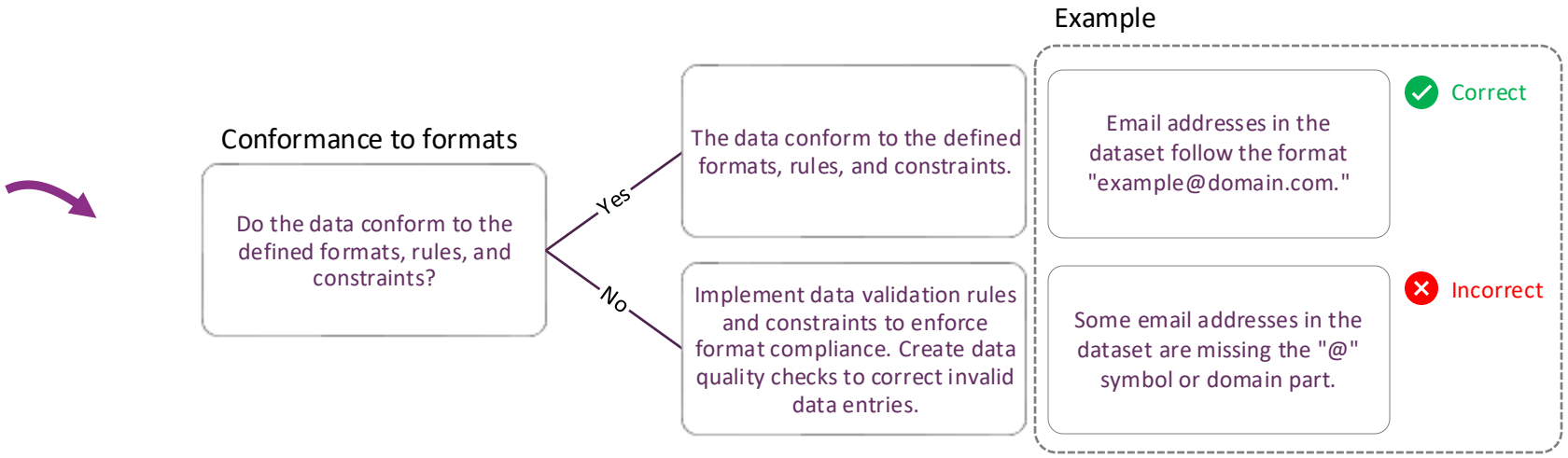
Example

- School water usage reports are uploaded to the SharePoint in near real-time by relevant schools. — Correct
- School water usage reports are only available in printed form and are irregularly sent by traditional mail. — Incorrect

Tool: Assessment on Validity

Data validation is the process of ensuring that data comply with predefined formats, rules, and constraints, thereby maintaining its accuracy, consistency, and reliability. This process involves checking data for errors, discrepancies, or invalid entries and applying corrective measures to uphold data quality standards.

- ❑ Use this tool to enforce data format compliance with validation rules, detect and correct errors using data quality tools, and ensure adherence to business rules through governance policies and rule-based engines.



Tool: Assessment on Uniqueness

Ensure datasets are free from duplicates and each record has a unique identifier.

Use de-duplication techniques to identify and remove duplicates, and implement unique identifiers, such as primary keys, to enforce record uniqueness and maintain data integrity.

- ❑ Use this tool to evaluate the presence of duplicate records and ensure unique identifiers are applied to each record and guiding necessary corrective actions.

No Duplicate Records

Are there any duplicate records in the dataset?

- Yes → There are duplicate records in the dataset.
- No → Identify and remove duplicate records using de-duplication tools and techniques. Implement unique identifiers to prevent duplication.

Example

- Correct: Each citizen is listed only once in the national registry.
- Incorrect: Some citizens are listed multiple times in the national registry.

Unique Identifiers

Does each record have a unique identifier?

- Yes → Each record has a unique identifier.
- No → Assign unique identifiers to each record. Use primary keys and constraints in databases to enforce uniqueness.

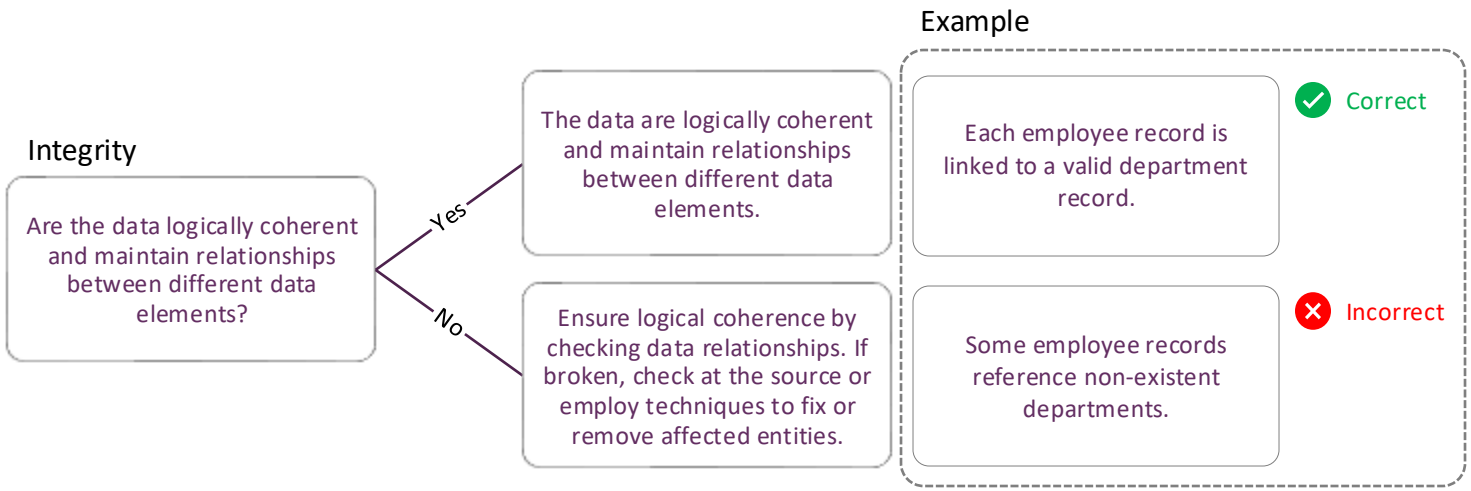
Example

- Correct: Each public contract has a unique contract ID.
- Incorrect: Some public contracts share the same contract ID.

Tool: Assessment on Integrity

Data integrity refers to the accuracy, consistency, and reliability of data, ensuring it remains logically coherent and maintains valid relationships between data elements throughout its lifecycle.

- Use this tool to ensure logical coherence by enforcing referential integrity, maintaining relationships between data elements, and validating data through foreign key constraints and regular audits.



You can jump to the desired page by clicking the links

Planning and timing: Weighting by value and effort

- Weighted Shortest Job First (WSJF) is a prioritisation method used to maximise value by focusing on use cases that provide the highest benefit in the shortest time.
- It calculates the ratio of Overall Value to Duration (Job size).
- Public administrations can use this to tool focus their effort on use cases with the highest WSJF score.

OVERALL VALUE
Impact of the outcomes in terms of their combined value

÷

DURATION (JOB SIZE)
Effort for completion

=

WEIGHTED SHORTEST JOB FIRST

This follows a proven methodology to ensure valuable projects by weighting value and effort

Inspiration is taken from https://scaledagileframework.com/wsjf/

Planning and timing: Weighting by value and effort

BUSINESS VALUE FOR PUBLIC ADMINISTRATION

- How much value is realised once this use case is solved?
- Do policymakers favour one use case over the other?
- Is there a potential penalty or other negative impact if a solution to this use case is delayed (e.g. reg. requirement)?

TIME CRITICALITY

- How urgent is this use case?
- How does the business value decay over time?
- Is there a fixed deadline? (recall a pilot on BDTI is a 6-month period)

RISK REDUCTION

- Does this use case reduce risk or create future opportunities?
- Will this use case open up new business opportunities?
- What else does this do for the public administration?

OVERALL VALUE

Impact of the outcomes in terms of their combined value

Courtesy of Accenture: "ML platform – Ideation Session"

Tool: Weighting by value and effort (part 1)

- To determine which use case to prioritise, a relative estimation between use cases is crucial.
- Estimating the relative value/size of a use case is done by assigning the appropriate value following the Fibonacci scale (1, 2, 3, 5, 8, 13, 21 and so on).
- Higher numbers are rarely used

Example using WSJF

Use cases	Business Value		Time Criticality		Risk Reduction		Overall Value
"Data insight through visuals"	5	+	5	+	1	=	11
"Timeseries forecasting"	8		8		8		24
…	…		…		…		…
…	…		…		…		…

We will revisit this later and estimate Duration and the final WSJF score

Inspiration is taken from https://scaledagileframework.com/wsjf/

Tool: Go/No Go Chart at end of Phase 1: Define

Category	*Topic*	*Go/No-Go Criteria*	*Status*
Strategic Alignment	Understanding of Strategic Fit	Does the use case align with strategic objectives?	Ready / Not Ready
Collaboration	Stakeholder Mapping and Engagement Skills	Are key stakeholders aligned and engaged?	Ready / Not Ready
	Communication Skills	Is communication clear across all teams?	Ready / Not Ready
Technical Skills	Data Management (SQL, ETL, Big Data tools)	Are data engineers and architects available?	Ready / Not Ready
	Programming (Python, R, JavaScript)	Are programmers skilled in required tools?	Ready / Not Ready
	Machine Learning and Statistical Analysis	Are data scientists capable of building ML models?	Ready / Not Ready
	Frontend Visualisation (Power BI, Tableau)	Are visualisation specialists available?	Ready / Not Ready
	DevOps and MLOps	Are DevOps and MLOps specialists in place?	Ready / Not Ready
	Cloud Architects	Are there Cloud Architects in the team?	Ready / Not Ready
Supporting Skills	Leadership and Project Management	Are strong leaders managing timelines and resources?	Ready / Not Ready
	Domain Expertise	Are SMEs engaged in the project?	Ready / Not Ready
	Policy and Compliance Understanding	Are data and policies compliant with regulations?	Ready / Not Ready
	Ethical Decision-Making	Are ethical standards being considered?	Ready / Not Ready
Data Quality	Data Quality Dimensions	Do data pass the quality checks in all dimensions?	Ready / Not Ready
Planning & Evaluation	Prioritisation (WSJF Scoring)	Have use cases been scored and prioritised?	Ready / Not Ready
	Metrics for Success	Are KPIs well-defined for evaluation?	Ready / Not Ready

For each criterion, mark as "Ready" or "Not Ready". All should ideally be marked "Ready" before moving to the next phase or there should be a valid reason for any not being ready

Phase 2: Design

Define | Design

This chapter will help you prepare for **designing** your pilot and find the right BDTI service offering.

Design refers to identifying analytics needs, selecting the appropriate BDTI service offering, assessing ethical considerations, and translating post analysis into actionable insights.

You can jump to the desired page by clicking the links

Getting familiar with the different analytics types

From information....

... to insights and actions

Descriptive	Diagnostic	Predictive	Prescriptive	Generative
What happened?	Why did it happen?	What will happen?	What should you do about it?	How should you execute?

Analytics

LLM & Generative AI

Machine Learning

Data Science

Courtesy of Accenture Data & AI Studio

Identifying analytics needs

Align with Objectives

Start by aligning the analytics category with the objectives. If understanding the past trends is crucial, **descriptive analytics** may be sufficient. If predicting future outcomes or optimising decisions, **predictive** or **prescriptive analytics** might be more appropriate

Data Availability and Quality

Consider the availability and quality of the data. Predictive and prescriptive analytics often require large, high-quality datasets, while **descriptive** and **diagnostic analytics** may be more forgiving of data limitations

Complexity of the Problem

Simple issues may be solved with descriptive or diagnostic analytics, while more complex challenges might require **predictive**, **prescriptive**, or even **generative analytics**

Desired Outcome

If the goal is to generate new content or solutions, generative analytics should be considered. For actionable recommendations, **prescriptive analytics** is likely the best fit

Tool: Data analytics types – Which direction?

When dealing with data analytics, numerous options are available to us. Selecting the appropriate method will aid in accomplishing your goal.

- ❑ Use the upcoming slides to delve into each of these disciplines

In which direction are you heading?

Diagnostic
Explaining why something happened

Descriptive
Focus on understanding what happened in the past

Predictive
Forecasting future outcome

Prescriptive
Suggest actions based on predictions

Generative
Generate novel content

Courtesy of Accenture Data & AI Studio: Identifying Data Analytic types

Tool: Data analytics types – Descriptive and Diagnostic

Descriptive Analytics

Focus on understanding what happened in the past. It provides insights into historical data to help understand trends, patterns, and behaviours

Types of analysis:
- Data aggregation and summarisation
- Reporting and dashboard creation
- Trend analysis over time

Use Case Suitability:
- Gain insight into past trends
- Understanding historical data to inform future decisions

Lisbon uses descriptive analytics for crowd monitoring.

Check the inventory of use cases to learn more about this and other applications.

Diagnostic Analytics

Explaining why something happened and identify the causes or factors that led to a specific outcome

Types of analysis:
- Root cause analysis
- Correlation and causality analysis
- Drill-down and data mining techniques

Use Case Suitability:
- Understanding patterns and identifying root causes
- Help uncover hidden patterns or relationships in data

Galicia uses diagnostic analytics to assess environmental impact in the mining sector.

Check the inventory of use cases to learn more about this and other applications.

Tool: Data analytics types – Predictive and Prescriptive

Predictive Analytics

Forecasting future outcomes based on historical data. Use statistical models and machine learning to predict trends and behaviours

Types of analysis:

- Predictive modelling and forecasting
- Risk assessment and scenario analysis
- Proactive decision making

Use Case Suitability:

- Predicting demand for services, forecasting budget requirements, or assessing the likelihood of policy success

Ljubljana uses predictive analytics for controlling stormwater flows.

Check the inventory of use cases to learn more about this and other applications.

Prescriptive Analytics

Not only predict outcomes but also suggests actions to achieve desired results. Provide recommendation based on predictive models with constrains/objectives

Types of analysis:

- Optimisation models and simulations
- Decision trees and recommendation systems
- Scenario planning with actionable insights

Use Case Suitability:

- Complex decision-making scenarios: Optimising resource allocation, planning emergency response strategies, improving delivery process

Tallin uses prescriptive analytics to improve urban transport planning.

Check the inventory of use cases to learn more about this and other applications.

Tool: Data analytics types – Generative

Generative Analytics

Focuses on creating novel content or scenarios based on existing data. It can be used to generate new text, images, simulations by learning from patterns in data

Types of analysis:

- Text and content generation
- Synthetic data creating for training AI models
- Scenario generation for planning and innovation

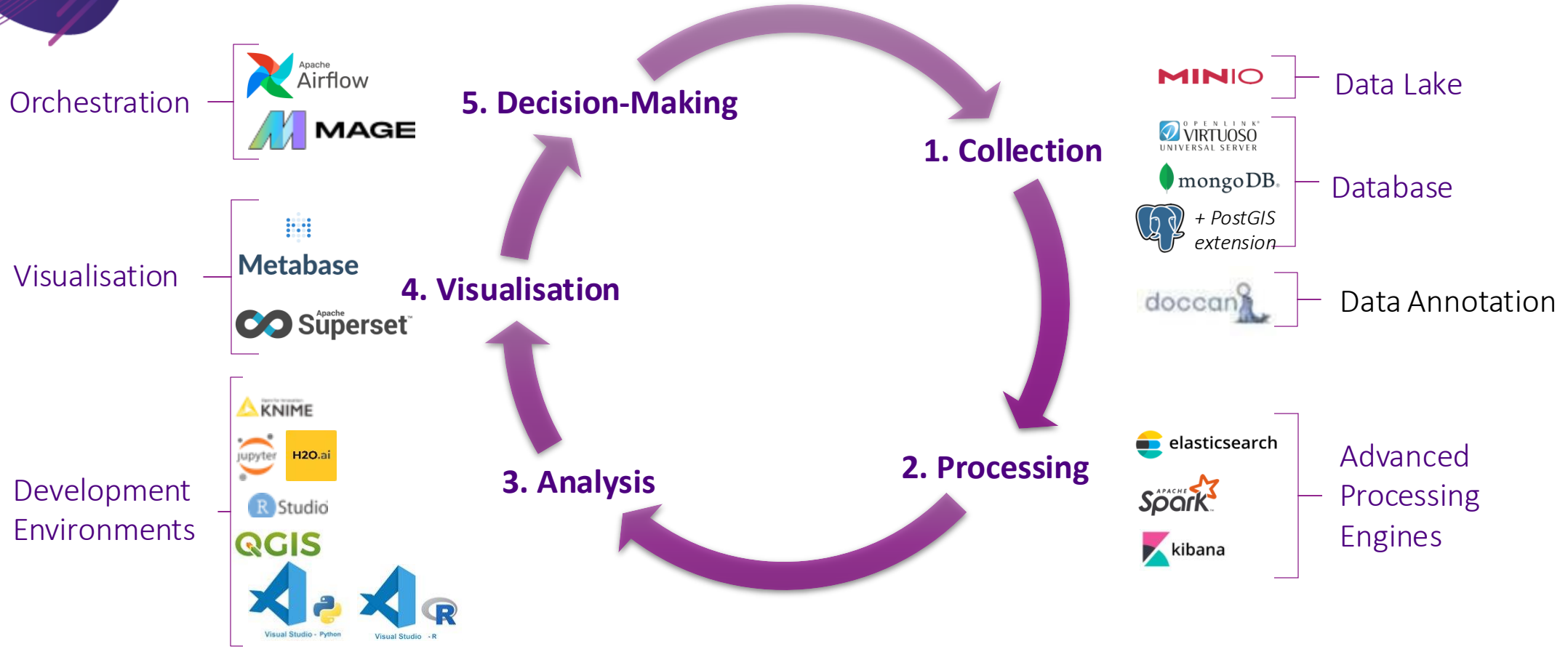
Use Case Suitability:

- Generating new policy drafts, creating training data for AI models, or simulate outcom in urban planning

Please note that advanced generative AI (e.g. LLM) is currently not supported on the BDTI platform

You can jump to the desired page by clicking the links

BDTI Service offerings in the analytics lifecycle



Tool: Linking to BDTI Service offering

BDTI has many service offerings. Selecting the appropriate service is crucial for accomplishing your feature/task.

Disclaimer: Listed tools are chosen from an ease-of-getting-started perspective. More advanced users might choose differently.

- ❑ Use this tool to quickly get started on the BDTI platform with the most appropriate tool

Alternative: Use the *"Linking analytic tools on BDTI"* Excel Spreadsheet

What type of analysis do we need for our pilot?

Descriptive Analytics

Tool Recommendation:
* *PostgreSQL* or *MongoDB* for querying historical data
* *Metabase* or *Apache Superset* for data visualisation and reporting

Next Question:
Do we need large-scale data processing?

- Yes → * Use *Apache Spark* for distributed processing
- No → * Continue with basic querying and visualisation tools

Diagnostic Analytics

Tool Recommendation:
* *KNIME* for root cause analysis and data mining
* *RStudio* for statistical analysis and correlation

Next question:
Are the data unstructured?

- Yes → * Use *MongoDB* for mapping flexible, unstructured datasets
- No → * Proceed with standard structured database solutions like *PostgreSQL*

Predictive Analytics

Tool Recommendation:
* *H20.ai* for machine learning model building
* *JupyterLab with Python* for custom predictive models

Next question:
Do we require real-time prediction?

- Yes → * Use *Apache Spark* or Elasticsearch for real-time analytics
- No → * Continue with batch processing tools like *H20.ai*

Prescriptive Analytics

Tool Recommendation:
* *KNIME* for decision modelling and workflow automation
* *RStudio* for optimisation and simulation models

Next question:
Are we handling complex workflows?

- Yes → * Use *Apache Airflow* or *Mage* to automate decision-making pipelines
- No → * Continue using *KNIME* for straightforward decision-making tasks

Generative Analytics

Tool Recommendation:
* *Doccano* for generating labelled data for NLP
* *JupyterLab* for generating custom content or exploring different scenarios

Tool: BDTI software stack vs. Skills – Data Lake

Tool Name	*Description*	*Difficulty to Learn*	*Difficulty Explanation*	*Learn More*
MinIO	MinIO is a high-performance, distributed object storage system designed for large-scale data infrastructure. It is compatible with Amazon S3 cloud storage service.	Moderate	MinIO requires understanding of storage concepts and setting up the server. While the initial setup is straightforward, configuring it for high availability, security, and performance optimisation can be challenging. Advanced features like erasure coding and object locking add to the complexity.	MinIO on BDTI's website MinIO's website MinIO's documentation

Tool: BDTI software stack vs. Skills – Databases

Tool Name	*Description*	*Difficulty to Learn*	*Difficulty Explanation*	*Learn More*
PostgreSQL	PostgreSQL is a powerful, open-source relational database management system (RDBMS) known for its robustness, extensibility, and SQL compliance.	Moderate	PostgreSQL involves learning SQL and relational database concepts. Basic CRUD operations are easy, but mastering advanced features like indexing, query optimisation, and stored procedures requires significant effort. Understanding its extensive configuration options and extensions also adds to the learning curve.	• PostgreSQL on BDTI's website • PostgreSQL's website • PostgreSQL's documentation
MongoDB	MongoDB is a popular NoSQL database that uses a flexible, JSON-like document structure, making it ideal for handling unstructured data and scaling horizontally.	Moderate	MongoDB requires understanding of document-oriented databases and basic CRUD operations. While it is easier to start with compared to relational databases, advanced features like aggregation pipelines, indexing strategies, and sharding for distributed data management add complexity.	• Mongo DB on BDTI's website • MongoDB's website • MongoDB's documentation
Virtuoso	Virtuoso is a multi-model database management system that supports relational, graph, and RDF data models, often used for linked data and semantic web applications.	Hard	Virtuoso supports multiple data models and query languages, including SQL, RDF, and SPARQL. This requires a deep understanding of both relational and semantic web technologies. The complexity is further increased by the need to manage and optimise performance across these different models.	• Virtuoso on BDTI's website • Virtuoso's website • Virtuoso's documentation

Tool: BDTI software stack vs. Skills - Orchestration

Tool Name	*Description*	*Difficulty to Learn*	*Difficulty Explanation*	*Learn More*
Airflow	Apache Airflow is an open-source platform to programmatically author, schedule, and monitor workflows, making it ideal for managing complex data pipelines.	Moderate	Airflow requires understanding of Python and workflow orchestration concepts. While setting up basic workflows is straightforward, managing complex dependencies, handling failures, and scaling the system for large workflows can be challenging. Advanced features like custom operators and sensors add to the complexity.	• Airflow on BDTI's website • Airflow's website • Airflow's documentation
Mage	Mage is an open-source data pipeline tool that simplifies the process of building, running, and monitoring data workflows, focusing on ease of use and scalability.	Moderate	Mage involves understanding data engineering concepts and programming skills. Setting up basic data pipelines is easy, but handling complex data transformations, integrations with various data sources, and ensuring data quality and reliability require more effort.	• Mage on BDTI's website • Mage's website • Mage's documentation

Tool: BDTI software stack vs. Skills – Environments (1/2)

Tool Name	*Description*	*Difficulty to Learn*	*Difficulty Explanation*	*Learn More*
JupyterLab	JupyterLab is an open-source web-based interactive development environment for Jupyter notebooks, code, and data, widely used in data science and research.	Easy	JupyterLab is easy to learn if you are familiar with Python and data analysis. Basic usage is straightforward, but advanced features like custom extensions, interactive widgets, and integrating with other tools and libraries require more effort. Managing large notebooks and optimising performance can also be challenging.	• JupyterLab on BDTI's website • JupyterLab's website • JupyterLab's documentation
RStudio	RStudio is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics, providing tools for data analysis and visualisation.	Easy	RStudio requires knowledge of the R programming language and statistical analysis. Basic usage is easy, but advanced features like package development, Shiny applications, and integrating with other tools and libraries add complexity. Understanding R's extensive ecosystem and optimising performance also require significant effort.	• RStudio on BDTI's website • RStudio's website • RStudio's documentation
VS Code	Visual Studio Code (VS Code) is a free, open-source code editor developed by Microsoft, known for its versatility, extensions, and support for various programming languages.	Easy	VS Code is user-friendly and easy to learn for basic coding and development tasks. Basic usage is straightforward, but advanced features like custom extensions, debugging, and integrating with other tools and services require more effort. Managing large projects and optimising performance can also be challenging.	• VS Code on BDTI's website • VS Code's website • VS Code's documentation

Tool: BDTI software stack vs. Skills – Environments (2/2)

Tool Name	*Description*	*Difficulty to Learn*	*Difficulty Explanation*	*Learn More*
KNIME	KNIME (Konstanz Information Miner) is an open-source data analytics, reporting, and integration platform, enabling users to create data workflows through a visual interface.	Moderate	KNIME requires understanding of data analytics concepts and some configuration skills. Basic usage is easy, but advanced features like custom nodes, complex workflows, and integrating with various data sources add complexity. Managing and optimising performance also require significant effort.	• KNIME on BDTI's website • KNIME's website • KNIME's documentation
H2O.ai	H2O.ai is an open-source platform for machine learning and artificial intelligence, offering tools for building and deploying predictive models.	Hard	H2O.ai requires knowledge of machine learning concepts, data preprocessing, and programming. The complexity increases with the need to optimise models, handle large-scale data, and integrate with other tools and libraries. Understanding H2O's extensive ecosystem and managing performance also require significant effort.	• H2O.ai on BDTI's website • H2O.ai's website • H2O.ai's documentation
QGIS	QGIS (Quantum GIS) is an open-source geographic information system (GIS) application that supports viewing, editing, and analysing geospatial data.	Moderate	QGIS requires understanding of GIS concepts and some configuration skills. Basic usage is easy, but advanced features like custom plugins, spatial analysis, and integrating with various data sources add complexity. Managing and optimising performance also require significant effort.	• QGIS on BDTI's website • QGIS's website • QGIS's documentation
doccano	doccano is an open-source annotation tool for text data, enabling users to create labelled datasets for natural language processing (NLP) tasks.	Easy	doccano is relatively easy to learn for basic text annotation tasks. Basic usage is straightforward, but advanced features like custom annotation schemes, integrating with other tools and services, and managing large annotation projects require more effort. Optimising performance and ensuring data quality can also be challenging.	• doccano on BDTI's website • doccano's website • doccano's documentation

From Data to Decision – creating data products

Data:

- Data is an abstract concept, often difficult to directly relate to our tasks. For it to be actionable, it requires **context**, **language**, and **expertise**.
- Data can be flawed—prone to bias, incompleteness, or errors.

Decision:

- Data aren't always 'evidence' or 'knowledge'. Not all data inform decisions.
- Improving methodology can enhance 'quality data,' influencing better information and decisions.
- High-quality data are crucial for accurate information, knowledge, and decisions, while poor quality data can lead to incorrect conclusions.

Source: Data Playbook, 'Making Data Useful, Useable and Shareable', p. 303, https://preparecenter.org/toolkit/data-playbook-toolkit-v1/

Data quality for OUTPUT data: FAIR principles

The FAIR guiding principles for scientific data management and stewardship offer a structured approach to evaluating various facets of data quality. This structure is based on four key areas: **findability**, **accessibility**, **interoperability**, and **reusability**, with clear criteria provided for each. It is advisable for data publishers to familiarise themselves with these FAIR guidelines prior to releasing data.

Findable	Accessible	Interoperable	Reusable
"Your data can be discovered by others"	"Your data can be made available to others"	"Your data can be integrated with other data"	"Your data can be reused by others"

Important: Data Quality Assessment – FAIR Principles

Findable

Unique Identifiers: Percentage of datasets with globally unique and persistent identifiers.
Metadata Richness: Average number of metadata elements per dataset.
Searchability: Number of datasets indexed in searchable repositories or registries.
Metadata Completeness: Percentage of datasets with complete metadata descriptions.

Interoperable

Standardised Formats: Percentage of datasets using standardised, machine-readable formats (e.g., JSON, XML, CSV).
Controlled Vocabularies: Percentage of datasets using controlled vocabularies or ontologies (e.g., Data Catalogue vocabulary (DCAT)) that follow FAIR principles.
Cross-Referencing: Percentage of datasets with qualified references to other datasets or metadata.
Semantic Annotation: Percentage of datasets with semantic annotations to enhance machine readability and interoperability.

Accessible

Protocol Standardisation: Percentage of datasets accessible via standardised communication protocols (e.g., HTTPS, FTP).
Open Access: Percentage of datasets that are openly accessible without restrictions.
Authentication and Authorisation: Percentage of datasets that support authentication and authorisation mechanisms where necessary.
Metadata Availability: Percentage of datasets with accessible metadata, even if the data themselves are restricted or no longer available.

Reusable

Licensing Clarity: Percentage of datasets with clear and accessible data usage licenses.
Provenance Information: Percentage of datasets with detailed provenance information (e.g., data source, creation date, modifications).
Community Standards Compliance: Percentage of datasets that comply with domain-relevant community standards.
Data Quality Indicators: Number of datasets with documented data quality indicators (e.g., accuracy, completeness, consistency).

IMPORTANT

Tool: Assessment on Data Quality – Findable

Ensure datasets are discoverable, accessible, and reusable by assigning unique identifiers (e.g., DOI) and creating rich metadata using standards like DCAT-AP.

Register datasets in searchable repositories such as the European Data Portal to enhance visibility and long-term access. Use this tool to verify compliance with metadata standards.

- ❑ Use this tool to ensure datasets have unique identifiers and rich metadata, registered in searchable repositories for accessibility and reuse.

Metadata and Documentation

- Does the dataset have a unique and persistent identifier (e.g., DOI, handle)?
  - Yes → The dataset has a unique and persistent identifier.
  - No → Obtain a unique and persistent identifier for the dataset by registering it with a data repository that provides DOIs, such as Zenodo or the European Data Portal.
- Is the dataset described with rich metadata, which is indexed in a searchable resource?
  - Yes → The dataset is described with rich metadata that are indexed in a searchable way.
  - No → Create detailed metadata using standards such as the DCAT-AP. Submit metadata to a searchable resource or repository, such as the European Data Portal or national open data portals.
- Is the dataset registered or indexed in a searchable resource?
  - Yes → Each record has a unique identifier.
  - No → Register the dataset in a searchable resource by submitting it to a data repository or catalogue that indexes datasets, such as the European Data Portal or Open Data portals of EU member states.

Data Identification

- Are the data and metadata provided with a clear and accessible data identifier?
  - Yes → The data and metadata are provided with a clear and accessible data identifier.
  - No → Assign a clear and accessible identifier to the data and metadata by using a persistent identifier system like DOI or Handle System.
- Are the identifiers used for the data globally unique and persistent?
  - Yes → The identifiers are globally unique and persistent.
  - No → Ensure the identifiers are globally unique and persistent by registering them with a recognised identifier service like CrossRef or DataCite.

Tool: Assessment on Data Quality – Accessible

Ensuring data access involves using open and standardised protocols for data retrieval, while clearly documenting access procedures for users.

Metadata should remain accessible even if the data are unavailable. Clear conditions for data access and reuse should be defined and communicated through appropriate licensing to promote transparency and compliance.

- ❑ Use this tool to ensure data are accessible via open protocols, retrieval procedures are documented, metadata remains available, and access conditions are clearly licensed.

Data Access

- Are the data retrievable by their identifier using a standardised communications protocol?
  - Yes → The data are retrievable by their identifier using a standardised communications protocol.
  - No → Implement a standardised communications protocol such as HTTPS or FTP for data retrieval. Ensure that the data can be accessed via a URL or API.
- Is the protocol open, free, and universally implementable?
  - Yes → The protocol is open, free, and universally implementable.
  - No → Use open and free protocols like HTTPS to ensure universal accessibility. Avoid proprietary protocols that require special software or licenses.
- Are the access procedures clearly described and accessible to users?
  - Yes → The access procedures are clearly described and accessible to users.
  - No → Document the access procedures in a README file or metadata record. Include instructions on how to access the data, any required software, and contact information for support.

Authorisation and Authentication

- Are metadata accessible even when the data are no longer available?
  - Yes → Metadata are accessible even when the data are no longer available.
  - No → Ensure that metadata are stored in a separate, persistent repository that remains accessible even if the data are removed. Use services like the European Data Portal or institutional repositories.
- Are there clear conditions for access and reuse of the data?
  - Yes → There are clear conditions for access and reuse of the data.
  - No → Define and communicate access and reuse conditions by applying a clear license to the data. Include this information in the metadata and documentation.

Tool: Assessment on Data Quality – Interoperable

Ensure data and metadata use standardised formats like XML, JSON, or RDF and adopt FAIR-aligned vocabularies such as EuroVoc or INSPIRE. Link datasets using persistent identifiers for context and follow metadata standards like DCAT-AP. Use common data formats like CSV or JSON to ensure interoperability with workflows.

- ❑ Use this tool to ensure data and metadata follow standards, adopt FAIR vocabularies, link datasets, and use interoperable formats.

Data and Metadata Standards

- Are the data and metadata represented using formal, accessible, shared, and broadly applicable language for knowledge representation?
  - Yes → The data and metadata represented are using the appropriate language for knowledge representation.
  - No → Use standardised formats and languages for data and metadata representation, such as XML, JSON, or RDF. Follow community standards like DCAT-AP for metadata.
- Are vocabularies used that follow FAIR principles?
  - Yes → There are vocabularies used that follow FAIR principles.
  - No → Adopt controlled vocabularies and ontologies that follow FAIR principles, such as the EuroVoc thesaurus or the INSPIRE data specifications. Reference these vocabularies in the metadata.
- Are the data and metadata linked to other data and metadata to provide context?
  - Yes → The data and metadata are linked to other data and metadata to provide context.
  - No → Create links between your data and related datasets or metadata records using persistent identifiers. Use standards like RDF to create these links.

Data Integration

- Are the data and metadata interoperable with applications or workflows for analysis, storage, and processing?
  - Yes → The data and metadata are interoperable with applications or workflows for analysis, storage, and processing.
  - No → Ensure interoperability by using common data formats (e.g., CSV, JSON) and metadata standards (e.g., DCAT-AP). Test your data with relevant tools and workflows to confirm compatibility.

Tool: Assessment on Data Quality – Reusable

Ensure data and metadata are richly described with accurate attributes, detailed methodologies, and clear context using standards like DDI.

Apply an accessible license, and document provenance with origins, processing steps, and transformations such as W3C PROV.

Comply with community standards like the INSPIRE Directive or DCAT-AP to enhance interoperability and usability.

- ❑ Use this tool to ensure rich metadata, apply a clear license, document provenance with W3C PROV, and meet standards like INSPIRE or DCAT-AP for interoperability.

Data Quality and Provenance

- Are the data and metadata richly described with a plurality of accurate and relevant attributes?
  - Yes → The data and metadata are richly described with the appropriate attributes.
  - No → Enhance the richness of your metadata by including detailed descriptions, methodologies, and context. Use standards like the Data Documentation Initiative (DDI) for social science data.
- Are the data released with a clear and accessible data usage license?
  - Yes → The data are released with a clear and accessible data usage license.
  - No → Apply a clear and accessible license to your data, such as a Creative Commons license. Include the license information in the metadata and documentation.
- Are the data associated with detailed provenance?
  - Yes → The data are associated with detailed provenance.
  - No → Document the provenance of your data by recording the origin, processing steps, and any transformations. Use standards like the W3C PROV for provenance information.

Community Standards

- Are the data and metadata compliant with relevant community standards?
  - Yes → The data and metadata are compliant with relevant community standards.
  - No → Identify and adhere to community standards relevant to public administration in the EU. For example, meet the INSPIRE Directive standards for spatial data or the DCAT-AP for data catalogues.

You can jump to the desired page by clicking the links

Working with data analytics and Ethical Considerations

When working with **open data**, **ethical practices** must be prioritised to ensure **responsible use**. By addressing these ethical considerations, open data can be **responsibly managed and shared**, fostering **trust and innovation** while ensuring **legal compliance**:

Privacy and Data Protection
Adhere to regulations like GDPR to protect personal data, ensuring they are anonymised or pseudonymised when shared.

Transparency
Ensure that data collection methods are clear and well-documented, promoting trust and accountability.

Data Accuracy
Verify that data are reliable and updated to avoid misinformation or misuse.

Inclusivity
Ensure open data are accessible to all, avoiding barriers that could exclude certain groups.

Intellectual Property Rights
Respect legal frameworks around data ownership and give proper attribution when necessary.

Tool: Questionnaire for Ethical Considerations (1/2)

How do we avoid bias and ensure fairness in analytic decision-making?

Topic	**Description**	***Question***
Bias	Identify Potential Sources of Bias. Bias refers to the systematic error introduced into data collection, analysis, or interpretation that skews results in a particular direction. This can occur due to various factors such as the selection of data sources, the design of algorithms, or the subjective decisions made by data scientists. Addressing bias is crucial to ensure that the data product provides fair, accurate, and representative insights for public policy and decision-making.	❑ *What biases might exist in the data being used?* ❑ *Are certain groups underrepresented or overrepresented?*
Fairness	Implement Fairness Mechanisms. Fairness mechanisms ensure that data collection, analysis, and usage are unbiased and equitable, promoting transparency and accountability. These mechanisms include implementing algorithms that prevent discrimination, ensuring diverse and representative data sets, and providing clear guidelines for ethical data usage. By doing so, public administration can make informed decisions that benefit all citizens fairly and justly.	❑ *What techniques can be used to ensure fairness, such as rebalancing datasets or applying fairness constraints in model training?*
Fairness	Engage Stakeholders in Fairness Discussions. Engaging stakeholders in fairness discussions involves actively involving diverse groups in conversations about equitable data practices. This includes organising forums and workshops to gather input from various community members, ensuring their perspectives are considered in decision-making processes. By fostering open dialogue, public administration can build trust and create data products that reflect the needs and values of all stakeholders.	❑ *How can we ensure that fairness considerations are transparent and understood by all stakeholders?*

Tool: Questionnaire for Ethical Considerations (2/2)

How do we avoid bias and ensure fairness in analytic decision-making?

Topic	**Description**	*Question*
Privacy	Assess Data Privacy Requirements. Assessing data privacy requirements involves identifying and understanding the legal and ethical standards for protecting personal information. This includes conducting thorough privacy impact assessments to evaluate potential risks and implementing robust data protection measures. By ensuring compliance with privacy regulations, public administration can safeguard citizens' data and maintain public trust.	❑ *What are the legal and regulatory requirements regarding data privacy?*
Privacy	Implement Data Anonymisation. Implementing data anonymisation involves transforming personal data into a format that prevents the identification of individuals. This process includes techniques such as data masking, aggregation, and encryption to ensure that sensitive information is protected. By anonymising data, public administration can use and share data responsibly while maintaining privacy and compliance with regulations.	❑ *What methods can be used to anonymise or pseudonymise data to protect individual privacy?*

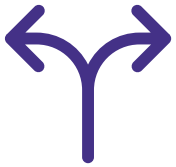
You can jump to the desired page by clicking the links

Post analysis translation into actionable insights

The ability to **interpret analytical findings** and **translate them into actionable insights** is crucial for any organisation. This process involves several key steps, each contributing to **transforming raw data** into **strategic decisions** that can drive success. This includes:

Interpreting Findings

Understanding what the data reveal about the problem or opportunity at hand.



Translating Insights into Strategic Recommendations

Developing specific action plans based on the analysis.

Communicating Insights to Stakeholders

Effectively conveying the insights to various stakeholder groups.

Tool: Translation into actionable insights

Interpreting Findings:

- What do the analytical results reveal about the problem or opportunity at hand?
- How do these findings align with the objectives defined at the beginning of the project?

Act

- ❑ Summarise the findings in a way that highlights the most relevant points
- ❑ Take inspiration from existing BDTI pilots : Pilot & Success Stories

Translating Insights into Strategic Recommendations:

- Based on the analysis, what specific actions should the organisation take?
- How do these recommendations support or enhance existing initiatives and programs?

Act

- ❑ Provide clear, evidence-based recommendations that are directly linked to the insights derived from the data
- ❑ Discover the Data in Publications Guide available at data.europa.eu

Communicating Insights to Stakeholders:

- How should the insights be communicated to different stakeholder groups, including decision-makers, operational teams, and the public?

Act

- ❑ Visualise data and insights where possible: example on BDTI
- ❑ Discover the Data Visualisation Guide available at data.europa.eu

Planning and timing: Weighting by value and effort

Following from our previous chapter on weighting by value and effort, we are now ready to finalise the estimation of the Duration (Job Size) and calculate the WSJF score.

DURATION (JOB SIZE)

Effort for completion

DURATION (JOB SIZE), is influenced by

- Effort for the use case analysis and delivery
- Potential technical risks
- Dependencies which could lead to blockers
- (Recall the pilot on BDTI is a 6-month period)

Inspiration is taken from https://scaledagileframework.com/wsjf/

Tool: Weighting by value and effort (part 2)

Continued example using WSJF

Use cases	Business Value	+	Time Criticality	+	Risk Reduction	=	Overall Value	/	Job Size	=	WSJF
"Data insight through visuals"	5		5		1		11		1		11.0
"Timeseries forecasting"	8		8		8		24		5		4.8
…	…		…		…		…		..		..
…	…		…		…		…		..		..

- ❑ Use this tool to focus effort on use cases with the highest WSJF score

Inspiration is taken from https://scaledagileframework.com/wsjf/

Tool: Go/No Go Chart at end of Phase 2: Design

Category	*Topic*	*Go/No-Go Criteria*	*Status*
Analytics Needs	Aligning objectives with analytics type (descriptive, predictive, etc.).	Have you identified the correct analytics type for your objective and ensured data quality supports it?	Ready / Not Ready
Data Quality	FAIR principles: Findable, Accessible, Interoperable, Reusable.	Are your datasets FAIR-compliant, with unique identifiers, rich metadata, and open access where possible?	Ready / Not Ready
Tool Knowledge	Familiarity with tools like KNIME, RStudio, PostgreSQL, JupyterLab, etc.	Do you or your team have the required expertise in the selected tools, or viable alternatives for quick upskilling?	Ready / Not Ready
Ethical Considerations	Addressing bias, privacy, data protection, and inclusivity.	Have you addressed potential biases, ensured fairness, and established compliance with data protection laws?	Ready / Not Ready
Translation to Insights	Interpreting findings, strategic recommendations, and stakeholder communication.	Can you interpret results clearly, provide actionable recommendations, and effectively communicate insights to stakeholders?	Ready / Not Ready
Weighting by Effort	Calculating WSJF scores (Value, Time, Risk Reduction, Effort).	Have you calculated WSJF scores to prioritise tasks effectively, focusing on the highest-value use cases?	Ready / Not Ready
Implementation Planning	Understanding dependencies, risks, and duration.	Have you evaluated technical risks, dependencies, and duration to ensure project feasibility within the timeline?	Ready / Not Ready

For each criterion, mark as "Ready" or "Not Ready". All should ideally be marked "Ready" before proceeding with the pilot

Conclusion

This document has been designed **to support public administrations in defining and designing their use cases** when considering **leveraging BDTI capabilities**. Throughout this journey, **public administrations** gain valuable **insights into adopting a data-driven approach**. The process encompasses several key steps (non-exhaustive):

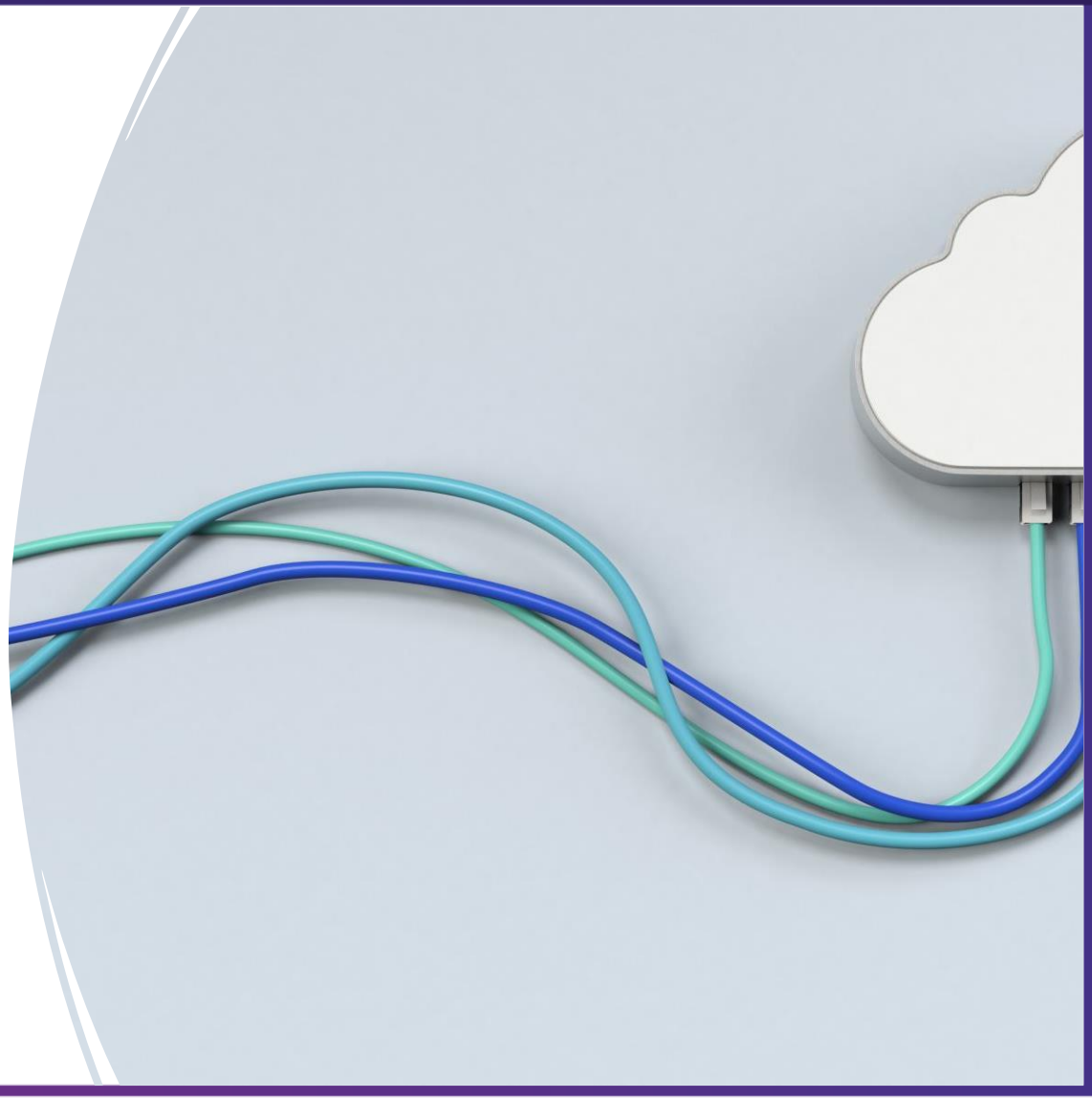
- Assessing opportunities and defining problems
- Aligning with the public administration's strategic vision
- Understanding the ecosystem and fostering collaboration
- Complying with the legal European framework
- Assessing technical and soft skills
- Conducting input and output data quality assessments
- Understanding various types of data analytics and identifying analytics needs
- Familiarising themselves with BDTI service offerings
- Considering ethical implications

This is a **dynamic document** that should be continuously revised based on the lessons learned and improvements suggested by stakeholders. By following this structured approach, public administrations can effectively leverage BDTI capabilities to achieve their strategic objectives and enhance their service delivery.

DIGITAL EUROPE PROGRAMME

Transform your data into actionable insights to improve public services.

Learn More



Appendix

References and Links

References and Links (1/4)

Data Knowledge Sources	
Six Sigma	https://www.sixsigma-institute.org/Six_Sigma_DMAIC_Process_Define_Phase_Six_Sigma_Project_Charter.php
CRISP-ML(Q) Lifecycle Process	https://ml-ops.org/content/crisp-ml
Data Playbook, 'Understanding how data matters', p. 69-70	https://preparecenter.org/toolkit/data-playbook-toolkit-v1/
Source: Data Playbook, 'Making Data Useful, Useable and Shareable', p. 303	https://preparecenter.org/toolkit/data-playbook-toolkit-v1/
FAIR Principles	https://www.nature.com/articles/sdata201618
WSJF	https://scaledagileframework.com/wsjf/

BDTI and Digital Europe Programme	
BDTI	https://big-data-test-infrastructure.ec.europa.eu/index_en
BDTI capabilities	https://code.europa.eu/bdti/bdti-user-documentation
Digital Europe Programme	https://digital-strategy.ec.europa.eu/en/activities/digital-programme
BDTI success stories	https://big-data-test-infrastructure.ec.europa.eu/pilots-success-stories/success-stories_en
BDTI Skills Studio	https://big-data-test-infrastructure.ec.europa.eu/resources/bdti-skills-studio_en
BDTI Example - Optimisation of Public Lighting	https://superset-1690295135.p1.bdti.dataplatform.tech.ec.europa.eu/superset/dashboard/2/?native_filters_key=Tfzg3YiXUQnCDwoeaZ7tpU-4NQnLAr_CGvq2tcSZy0rb93LfhnjKC3Ye_fn_qvf2

References and Links (2/4)

Legal Framework:	
Open Data Directive (2019)	https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024
European Data Strategy (2020)	https://digital-strategy.ec.europa.eu/en/policies/strategy-data
Data Governance Act (2020)	https://digital-strategy.ec.europa.eu/en/policies/data-governance-act
Data Act (2022)	https://europa.eu/!GMcVf8
Interoperable Europe Act (2024)	https://europa.eu/!n3cbmF
AI Act (2024)	https://europa.eu/!Yh74XM
Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information	Summary - https://eur-lex.europa.eu/EN/legal-content/summary/open-data-and-the-reuse-of-public-sector-information.html Full text - https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019L1024
EU Data Protection Rules	https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en

Inspiration for intergovernmental Agreements:	
Helsinki Region Infoshare (HRI)	https://hri.fi/en_gb/hri-service/what-is-hri/
Ghent and Bruges Local Open Data Economy	https://stad.gent/en/city-governance-organisation/city-policy/ghent-international/funded-projects/project-lode-local-open-data-economy
Smart Dublin Initiative	https://smartdublin.ie/about/

References and Links (3/4)

Guidelines and data protection rules	
Guidelines and templates for agreements between public administrations	https://interoperable-europe.ec.europa.eu/sites/default/files/document/2017-10/Guidelines and Templates for Collaboration Agreements.pdf
Data Visualisation Guide	https://data.europa.eu/apps/data-visualisation-guide/
Data in Publications Guide	https://data.europa.eu/apps/data-in-publications-guide/

BDTI Software Stack vs. Skills	
MinIO	MinIO on BDTI's website MinIO's website MinIO's documentation
Airflow	Airflow on BDTI's website Airflow's website Airflow's documentation
Mage	Mage on BDTI's website Mage's website Mage's documentation
JupyterLab	JupyterLab on BDTI's website JupyterLab's website JupyterLab's documentation
RStudio	RStudio on BDTI's website RStudio's website RStudio's documentation

References and Links (4/4)

VS Code	VS Code on BDTI's website VS Code's website VS Code's documentation
Knime	KNIME on BDTI's website KNIME's website KNIME's documentation
H2O.ai	H2O.ai on BDTI's website H2O.ai's website H2O.ai's documentation
QGIS	QGIS on BDTI's website QGIS's website QGIS's documentation
doccano	doccano on BDTI's website doccano's website doccano's documentation