Open Data Toolkit

SDAIA

March 2023

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Open data toolkit – Intent and target audience

The Open Data Toolkit is designed to help government entities, publishers, and users understand the basic concepts of Open Data including how to plan and implement an open government data program by providing toolkits on:

- Developing an open data inventory
- Assessing value and determining high-value datasets
- Developing open data publishing plan
- Overview of global use-cases for open data

This toolkit provides a comprehensive step by step approach to help entities release more open data in a simplified manner while encouraging the development of impactful use-cases that unlock value for the Kingdom

Below are some of the most important target audience groups that could make use of this handbook:

- Those concerned with managing open data initiatives and activities in the government sector;
- Government sector chief data officers;
- SMEs and entrepreneurs
- Smart application developers;
- Researchers in academic and other institutions; and
- Those interested in government data practices outside the government sector.

Toolkit overview

Data inventory set-up guide

Data value assessment toolkit

Dataset development toolkit

Publishing plan toolkit

Use cases of Open Data around the world

Back to homepage

Table of contents

1 **Data inventory toolkit**

2 **Data value assessment toolkit**

3 **Dataset development toolkit**

4 **Publishing plan toolkit**

5 **Use cases of Open Data around the world**

What is a data inventory?

A **data inventory** (sometimes referred to as a **data map or data mapping**), is a **comprehensive catalog** of **data assets** held by an organization. A well-maintained data inventory includes **up-to-date and detailed information** regarding the data, as well as the **source of the data** within the organization

Data inventories can be built leveraging a 5-step approach

1 Establish user needs and requirements

1.A Understand users, their work and its context as the system should support achieving goals
1.B Based on needs identified, produce a set of requirements; statements that specify what an intended product should do, or how it should perform (functional vs non-functional requirements)

2 Collect data to be stored

2.A Communicate with relevant departments and issue requests to collect desired datasets to be stored within the inventory

3 Select inventory software

3.A Based on needs and requirements of the inventory, entities must assess and select a relevant software platform

- **Open-source software** (free to download, requires expertise to implement and maintain)
- **Commercial software** (paid software, software vendor owns, creates and maintains the source code
- **Software service** (software vendor owns and distributes a software platform, or also hosts and manages the data)

4 Assign roles and responsibilities

4.A Establish governance structure to maintain and update the data inventory while detailing communication structures between departments
4.B Assign roles and responsibilities including update timeline, request issuance

5 Manage the open data inventory

5.A Ensure adherence to the FAIR principle; data must be **F**indable (accurate metadata); **A**ccessible (open, free, and universally implementable); **I**nteroperable (applicable language for knowledge representation); and **R**e-usable (meet community needs)
5.B Set data curation and preservation principles which combine policies, strategies and actions to ensure the most accurate rendering possible of the data over time

As a first step, it is helpful to lay down the foundations of the data inventory by answering the following questions

Sample questions to create a data inventory

1 *What data is currently being collected?*

2 *Who is the data owner?*

3 *How is it being collected (e.g., internally; externally; manual; automatic; raw; aggregated)*

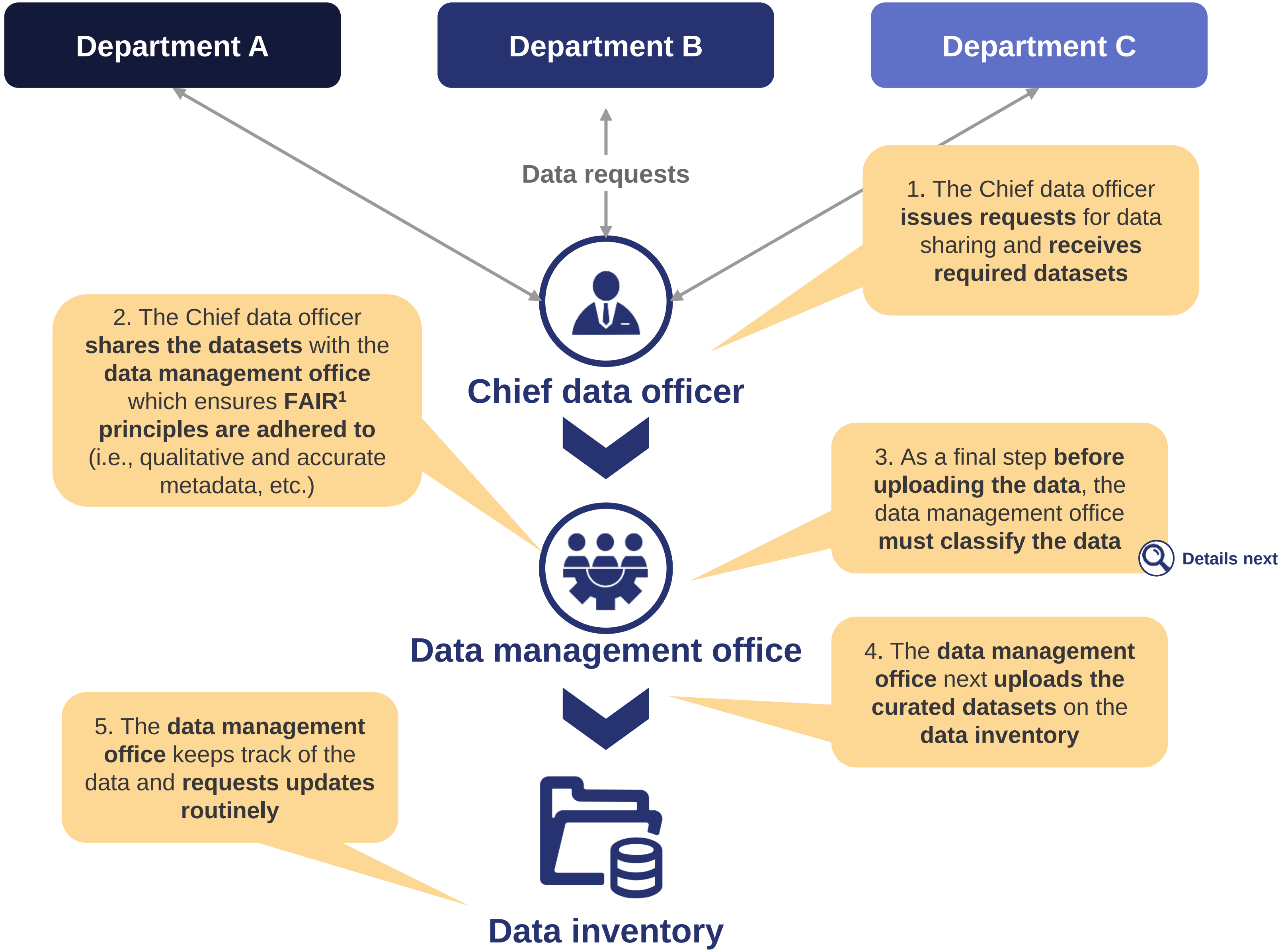
4 *In which format?*

5 *Where it is being stored? Is it database-structured or individual files?*

6 *Can it have any value for any external user (e.g., data might not be relevant to users)*

7 …

When first building a data inventory it is important to first gather the required datasets



Department A

Department B

Department C

Data requests

Chief data officer

Data management office

Data inventory

1. The Chief data officer **issues requests** for data sharing and **receives required datasets**

2. The Chief data officer **shares the datasets** with the **data management office** which ensures **FAIR**[1] **principles are adhered to** (i.e., qualitative and accurate metadata, etc.)

3. As a final step **before uploading the data**, the data management office **must classify the data**

Details next

4. The **data management office** next **uploads the curated datasets** on the **data inventory**

5. The **data management office** keeps track of the data and **requests updates routinely**

Notes: 1. Data must be **F**indable (accurate metadata); **A**ccessible (open, free, and universally implementable); **I**nteroperable (applicable language for knowledge representation); and **R**e-usable (meet community needs)

Before uploading to the inventory, data must be classified according to the interim data classification published by NDMO

Source: https://sdaia.gov.sa/ndmo/Files/PoliciesEn.pdf

Public

- Data shall be classified as “Public”, if unauthorized access to has no impact on
  - National Interest,
  - Organizations,
  - Individuals,
  - Environment

Confidential

- Data shall be classified as “Confidential”, if unauthorized access to causes
  - Negative effect on government
  - Damage to entity assets
  - Damage of nature

Secret

- Data shall be classified as “Secret”, if unauthorized access to affections
  - National interest
  - Financial loss to KSA orgs
  - Harm or injury to life

Top Secret

- Data shall be classified as “Top Secret”, if unauthorized access to violates
  - Conventions, treatise of national interest
  - KSA org functionality
  - Environment

> “ **Data is open by default**

NDMO
مكتب إدارة البيانات الوطنية
National Data Management Office

When requesting data from different entities, the data management office should set a standardized format for receiving the data

Data Inventory Template

Click for sample inventory template

ID	Name of dataset	Resp. dept.	Resp. person	Description	List of fields	Format	Expect publishing date	Timeline	Format	Potential data users
1.1.1	**Spend on IT infra.**	Infra. Deputyship	TBD	Total value of spending on technology infrastructure	Infra, Spend in SAR, Region	Raw	December 31st, 2023	Yearly	xls	IT businesses
1.1.2	**Number of tech start-ups**	Entrepreneurship deputyship	TBD	Total number of technology startups in the Kingdom	Startup name, Field, HQ, region	Statistics	December 31st, 2023	Yearly	xls	Investors, government
2.1.3	**Ministry budget**	Strategy department	TBD	Total allocated budget for the ministry	Budget, Year, Actual, Projected	Aggregated	January 1st , 2023	Yearly	xls	Citizens
2.2.4	**Total value of tech exports**	Technology market development	TBD	Total value of technology exports and re-exports form the Kingdom	Tech, Value exported, Year	Statistics	December 31st, 2023	Yearly	xls	Technology players
3.1.1	**Value of export split by export country**	Technology market development	TBD	Split of exports value by trading partner country on annual basis	Country, Value of export, Tech, Year	Statistics	December 31st, 2023	Yearly	xls	Technology players

Table of content

1 **Data inventory toolkit**

2 **Data value assessment toolkit**

3 **Dataset development toolkit**

4 **Publishing plan toolkit**

5 **Use cases of Open Data around the world**

Open high-value datasets must adhere to technical and measurement characteristics

I. Technical requirements

Openly published and free-of-charge

Extensive and comprehensive metadata

Machine readable format

II. Ease of measurement and update requirement

Measured accurately and qualitatively

Refreshed routinely with ease

Efficient measurement processes

Sources: OECD; Worldbank; European Commission; W3.org; Kearney analysis

There are 4 main approaches to measure the value of datasets, each with specific benefits and shortfalls

	Cost to value approach	Market value approach	Economic value approach	Stakeholder approach
	This method is based on the **cost to produce and store data**, as well as the cost to **replace lost data** and what the **impact on cash flow** would be	This approach is based on what **others pay for comparable data on the open market**, by observing those selling data (thus drawing on an example of data value) and **calculating the data selling price**	This approach is based on **impact of data on the GDP, specific industries** or valuation of specific **businesses built** on top of these **datasets**	This approach is based on the **demand for specific datasets** and the overall **value for the** society
Pros	*Easy to measure*	*Easy to measure*	*Ties to actual benefit to the economy / society*	*May be measured by multiple parameters (i.e., not limited to monetary value only)*
Cons	Undervalues data because as it ignores the question of how does data becomes use cases	Lack of basis for fair market price for unique datasets	Hard to measure, some data might not have clear economic benefit but is of high importance	Hard to compare value between different datasets that have different metrics
Examples	*Manually collected data, where volume directly correlates with cost (e.g., human genome dataset – USD 3 Bn)*	*Comparing similar paid market research data provided by different companies*	*Lensa AI got USD 16 Mn revenue in 2022 for the "magic avatar" which allows users to download photos of themselves in different settings*	*Competitors within a given industry requesting market insights regarding supply chains*

To assess whether a datasets meets identified requirements, a scoring system is developed (1/3)

A Prerequisite assessment

Is the information a dataset or statistics?

Question	*Points*
A.1 What is the format of the information?	
– Text / PDF	0
– Tabular (xls / CSV / other)	1
A.2 Does the file contain raw or aggregated information?	
– Aggregated	0
– Raw	1
A.3 How many rows are represented in the file?	
– <100	0
– 100 – 500	1
– >500	2
A.4 What is the timestamp for the recorded data?	
– Yearly	0
– Monthly	1
– Daily or more frequent	2

B Cost-model assessment

How costly is it to assemble the dataset?

Question	*Points*
B.1 How many men-hours required to replicate the dataset?	
– <1 day	0
– 1 – 5 days	1
– >5 days	2
B.2 What is the required investment to collect this data?	
– <1K USD	0
– 1K – 1Mn USD	1
– >1Mn USD	2
B.3 If data is being completely lost, is it possible to reassemble same dataset?	
– Yes	0
– Partly	1
– No	2

To assess whether a datasets meets identified requirements, a scoring system is developed (2/3)

C Comparable pricing assessment

Is the dataset highly priced on the market?

Question	*Points*
C.1 Is comparable data available on the market?	
– Yes	0
– No	1
C.2 What is the price of the dataset on the market?	
– <500 USD	0
– 500 – 1000 USD	1
– >1000 USD	2

D Economic value assessment

Does the dataset have potential to generate economic value?

Question	*Points*
D.1 Can the dataset be leveraged to generate innovative services?[1]	
– No potential	0
– Add-on service	1
– New, innovative service	2
D.2 Does the dataset enable operational efficiency?	
– No cost benefits	0
– Eliminates duplication of efforts	1
– Enables cost savings and eliminates duplication of efforts	2
D.3 Does the dataset have the potential to create jobs?	
– No potential	0
– Creation of new departments within companies	1
– Creation of new companies within the industry	2

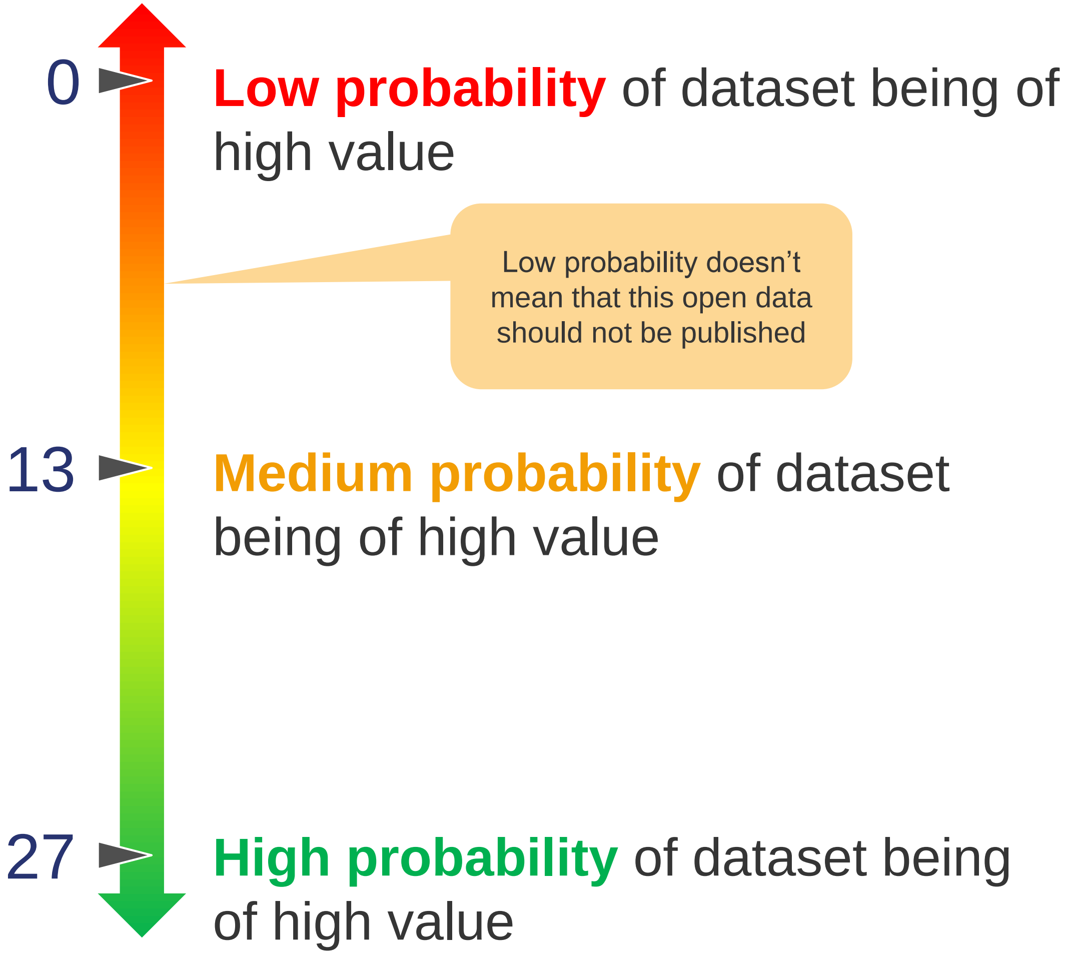
Notes: Examples include Berlin public transport planning system which optimizes routes based on real-time information about all buses and metro at the moment

To assess whether a datasets meets identified requirements, a scoring system is developed (3/3)

E Demand assessment

Is the dataset in high demand?

Question	*Points*
E.1 How many private sector players requested the dataset?	
– 0	0
– 1 – 10	1
– >10	2
E.2 How many public sector players requested the dataset?	
– 0	0
– 1 – 5	1
– >5	2
E.3 Is there a trend around the services relevant to the dataset?	
– No	0
– Regional	1
– Global	2



Low probability of dataset being of high value

Low probability doesn’t mean that this open data should not be published

Medium probability of dataset being of high value

High probability of dataset being of high value

Table of contents

There are 7 main steps when first building a dataset

1. **Identify source and mining methodology**
2. **Identify required fields**
3. **Identify roles and responsibilities for data collection**
4. **Build DB architecture, connections to existing data, master data management, etc.**
5. **Initiate data collection and storing**
6. **Clean data and run a quality check**
7. **Schedule regular updates**

1
2
3
4
5
6
7

1. Based on the data you want to collect, define where and how you can get it

Non-exhaustive

	Method	When to use	How to collect data
Primary data collection	**Direct instrumental measurement**	When data can be easily measured (e.g., daily temperature, gas consumption)	Put meters and record data from the meters on a regular basis
	Survey	To understand the general characteristics or opinions of a group of people	Distribute a list of questions to a sample online, in person or over-the-phone
	Interview/focus group	To gain an in-depth understanding of perceptions or opinions on a topic	Verbally ask participants open-ended questions in individual interviews or focus group discussions
	Online user inputs	When data is available for online activities (e.g., usage of online government services)	Add code elements on website to record user statistics
Secondary data collection	**Meta research**	When multiple similar datasets are available for a target topic	Find existing datasets that have already been collected and align them between each other
	Digitalization of analog records	When individual data pieces are available in analogue (e.g., paper) format	Access manuscripts, documents or records from libraries, depositories or the internet and convert them to digital records

2. Based on the data you want to collect, identify required fields that will comprise the dataset

Sample guiding questions:

Non-exhaustive

1 **What are the main drivers of the dataset?**
(e.g., for technology exports drivers may include the volume, the destination country; etc.)

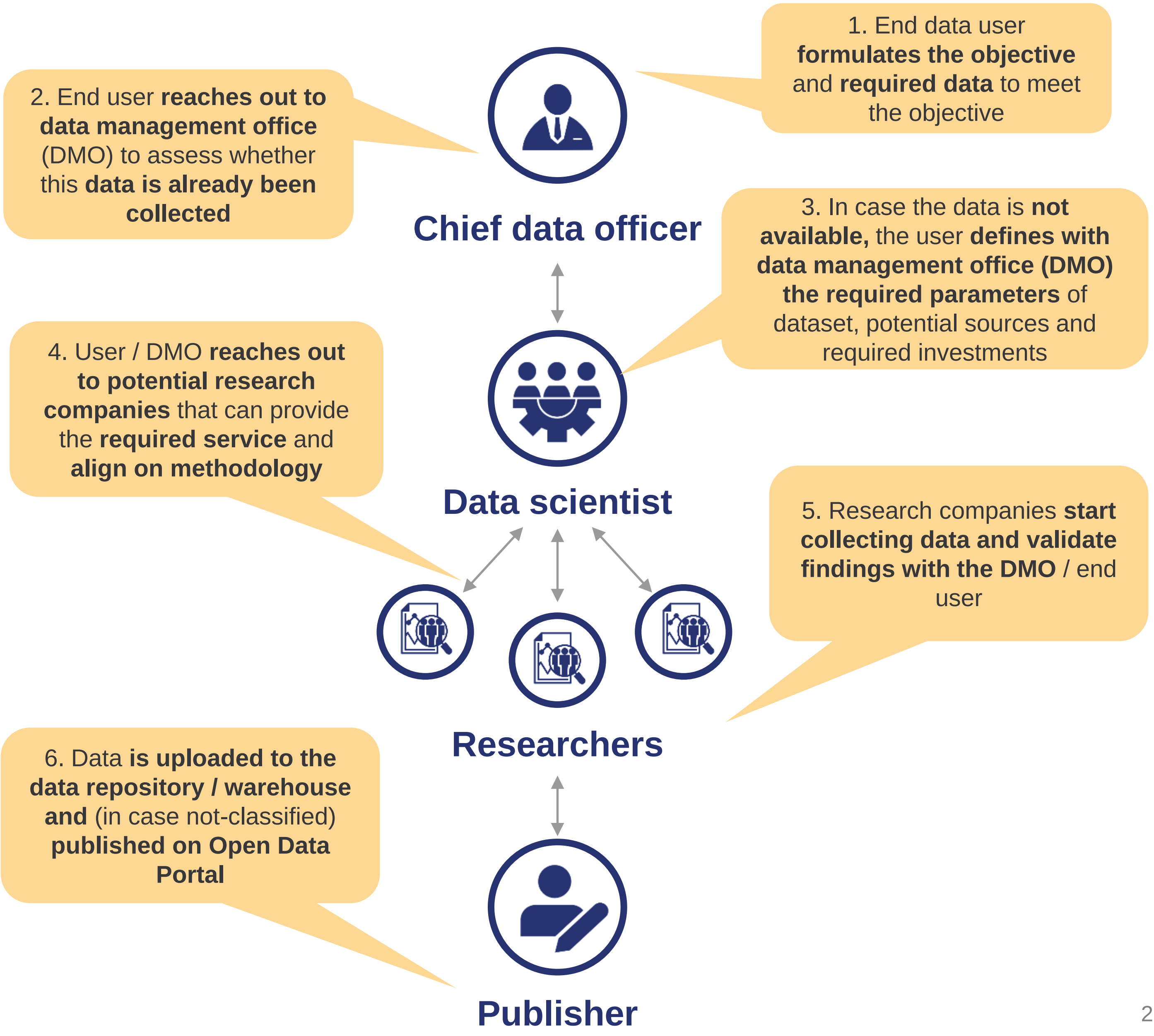
2 **What are the relevant fields listed across leading countries?**
(e.g., for technology exports relevant fields on the Irish data portal include destination country; value of exports; etc.)

3 **What are the specific stakeholder inputs?**
(e.g., for technology exports senior executives would like to understand what types of technologies are exported; etc.)

4 …

3. Identify the roles and responsibilities of individuals across the data management office to ensure collection of the required data

Non-exhaustive



1. End data user **formulates the objective** and **required data** to meet the objective

2. End user **reaches out to data management office** (DMO) to assess whether this **data is already been collected**

3. In case the data is **not available,** the user **defines with data management office (DMO) the required parameters** of dataset, potential sources and required investments

4. User / DMO **reaches out to potential research companies** that can provide the **required service** and **align on methodology**

5. Research companies **start collecting data and validate findings with the DMO** / end user

6. Data **is uploaded to the data repository / warehouse and** (in case not-classified) **published on Open Data Portal**

1 2 3 4 5 6 7

4. Build DB architecture, connections to existing data, master data management, etc.

1 **Determine the purpose of your database**
This helps prepare you for the remaining steps

2 **Find and organize the information required**
Gather all of the types of information you might want to record

3 **Divide the information into tables**
Divide your information items into major entities or subjects

4 **Turn information items into columns**
Decide what information you want to store in each table. Each item becomes a field and is displayed as a column in the table.

5 **Specify primary keys**
Choose each table’s primary key. The primary key is a column that is used to uniquely identify each row (e.g., product ID)

6 **Set up the table relationships**
Add fields to tables or create new tables to clarify the relationships, as necessary

7 **Refine your design**
Analyze your design for errors. Create the tables and add a few records of sample data. Make adjustments as needed

8 **Apply the normalization rules**
Apply the data normalization rules to see if your tables are structured correctly. Make adjustments to the tables, as needed.

Source: Microsoft

5. Implement chosen methods to measure or observe the variables required to develop the dataset Best practices include:

1 **Record all relevant information as soon as it is obtained**
(e.g., temperature fluctuations in specific regions in KSA)

2 **Sanity check and review any manual entries for errors?**
(e.g., temperature fluctuates by 100 degrees should raise questions as to a probable error resulting from an extra 0)

3 **Assess the reliability and validity to get an indication of the data quality**
(e.g., compare temperature fluctuations to historical data and neighboring regions to sanity check reported numbers)

4 **Store data in a readable and accessible manner for future reference**
(e.g., include data in a spreadsheet with exact timings of entries along with the corresponding region)

1 2 3 4 5 6 7

6. The main tasks to be carried out when cleaning data include:

1. **Omitting unwanted observations**
   Remove observations that are irrelevant

2. **Unifying the data structure**
   Ensure data from different sources is consistent and in a unified structure

3. **Standardizing the data**
   Ensure data collected uses the same units of measurement

4. **Removing outliers**
   Remove one-off data that ay skew findings / ensure normalization

5. **Fixing cross-set data errors**
   Ensure data from different sources do not contradict each other

6. **Resolving syntax errors**
   Remove whitespace, check for spelling mistakes, etc.

7. **Dealing with missing data**
   Remove associated entries, develop assumptions for missing values, etc.

8. **Validating the data**
   Ensure that all steps have been caried out properly

1
2
3
4
5
6
7

6. What does a machine readable, clean dataset look like?

Examples of machine readable, clean data vs Examples of data that is not machine readable or clean

Non-exhaustive

Machine readable, clean data

Date	Age	Gender	Postcode
20/10/2023	12	Male	2580
10/01/2023	40	Female	1462
02/11/2022	28	Male	3476
12/05/2022	33	Female	0987
19/01/2023	57	Female	1190

Corrupted / unstructured / incomplete data

Date	Age	Gender	Postcode
20/10/2023	Twelve	Male	Riyadh
10/01/2023	40	Woman	1462
02/11/2022	Twenty-eight	M	Tahlia
Tuesday 12, March	330	Fem	
19/01/2023	xx	Female	1190

7. Schedule regular updates

Final considerations to keep in mind…

1 **Keep a schedule of how often the datasets are to be updated**
Assess update frequency, pre-determine refresh dates, and track updates

2 **Assign responsibilities**
Once refresh timelines for datasets have been determined assign respective resources to track and update according to pre-determined dates

3 **Store historical data**
It is important to store historical data before updating in-case of any loss in content or error in reporting

Example

For petrol prices, it wouldn't make sense to only update the dataset once a month as interested users would likely look somewhere else for fresher data

Back to homepage

Table of contents

1 **Data inventory toolkit**

2 **Data value assessment toolkit**

3 **Dataset development toolkit**

4 **Publishing plan toolkit**

5 **Use cases of Open Data around the world**

Data inventories may be converted into publishing plan in 7 steps

1 Classify the data at hand

2 Ensure no existing IP liabilities

3 Define publishing date

4 Identify update requirement and timeline

5 Identify responsible departments and assign roles and responsibilities

6 Set-up internal processes

7 Publish dataset

Illustrative
Non-exhaustive

Publishers need to track published datasets and ensure they are routinely refreshed

Publishing template

Click for sample publishing template

ID	Name of dataset	Resp. dept.	Resp. person	Description	Date published	Refresh date	Update frequency
1.1.1	**Spend on IT infra.**	Infra. Deputyship	Omar Assiri	Total value of spending on technology infrastructure	December 31st, 2023	December 31st, 2024; etc.	Yearly
1.1.2	**Number of tech start-ups**	Entrepreneurship deputyship	Ahmad Khalil	Total number of technology startups in the Kingdom	December 31st, 2023	March 30th, 2024; June 30th, 2024; September 20th, 2024; etc.	Quarterly
2.1.3	**Ministry budget**	Strategy department	Omar Assiri	Total allocated budget for the ministry	January 1st , 2023	January 1st, 2024; etc.	Yearly
2.2.4	**Total value of tech exports**	Technology market development	Fahd Taweel	Total value of technology exports and re-exports form the Kingdom	December 31st, 2023	June 30th, 2024; December 31st , 2024; etc.	Semi-annually
3.1.1	**Value of export split by export country**	Technology market development	Fahd Taweel	Split of exports value by trading partner country on annual basis	December 31st, 2023	June 30th, 2024; December 31st , 2024; etc.	Semi-annaully

Back to homepage

Table of contents

1 **Data inventory toolkit**

2 **Data value assessment toolkit**

3 **Dataset development toolkit**

4 **Publishing plan toolkit**

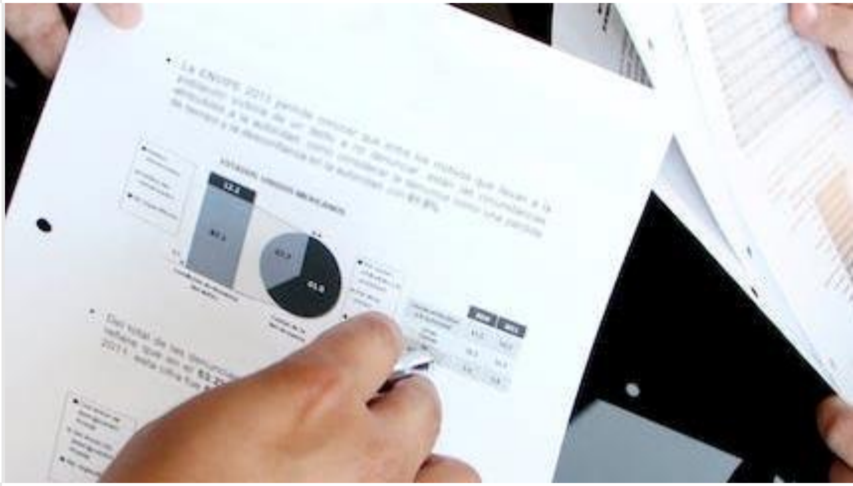
5 **Use cases of Open Data around the world**

Use cases for open data are spread across 5 main dimensions...

1. Government accountability, transparency, and policy effectiveness assessment

2. Effective government services and data reuse

3. Informed decision making, increasing potential value generation



4. Eased ability to conduct research

5. Increased AI training and innovation

Sample global use case

Many countries leverage open data to develop innovative use-cases for the education sector

Non-exhaustive

Informed educational choices for parents and children

Information such as list of schools, quality ratings, demography of students, subjects and majors offered allows better decision-making.

Example: QEdu, in Brazil, displays public information on quality of learning in schools – which benefits both parents and education managers

Increased transparency

Increased transparency on policies, funding and resources ensures accountability

Example: Afla.md in Moldova publishes data on planned expenses for all schools, contributing in transparency to public spending

Support public policy decision

Open data is being used to ensure access to quality education for students

Example: Mexico government is leveraging open data from multiple sources to analyze demand for educational facilities and plan construction

Open Science and research

Greater access to data and visibility to research enhances transparency and efficiency

Example: Netherlands Society for Biomaterials and Tissue Engineering is publishing data on links for material-specific topographical and chemical properties to gene expression database available for researchers across the world

1. Government accountability, transparency, and policy effectiveness assessment

Open data and transparency around the world

Overview

- Data from **26 countries**, shows **significant effect of data transparency reforms on government bond spreads**, (difference between the interest rate on a US government bond and that on a bond issued by another country). It is **used as a measure of a country's risk** when it comes to **investing`**
- In Africa, **13 countries** have **implemented the enhanced General Data Dissemination Standard** including Nigeria, Senegal, Sierra Leone, and Tanzania. In the **Asia-Pacific region**, Bhutan, Nepal, and Samoa, have **also implemented it**
- These countries **publish key economic data**, such as real **GDP growth through a National Summary Data Page**, which **provides policy makers**, **investors**, **rating agencies**, and **the public** with **easy access to information** critical for monitoring economic conditions

Benefits

The General Data Dissemination Standard suggests that it **improves coordination between a country's central bank**, **ministry of finance** and **statistics institute**, the three institutions involved in data dissemination. This enhanced coordination represents **an improvement in governance and overall decreased country risk premium**

2. Effective government services and data reuse

Admin effectiveness in Norway

Overview

- Statsregnskapet is a website that **visualizes government spending** and **budgets** on the basis of **publicly available government statistics**
- In doing so, it aims to **create transparency in government spending**
- Statsregnskapet **utilizes the data published by the Directorate for Public Administration and Financial Management (DFØ),** Norway's national body for public sector finances
- DFØ **publishes updated accounting figures** for the state **every month**. The detailed accounting information throughout the year and at the end of the year is the basis for the updated state accounts, which are **routinely published and made openly available**

Benefits

The website visualizes **where the money is coming from**; what the money **was spent on**; which **departments spent the money**; and what **the monthly and yearly developments are**. This allows for **reduced burdens on administrative staff** while **minimizing risk of errors and theft**. Data may be reused to **forecast spending** and **analyze trends** to determine **sectors of interest**

2. Effective government services and data reuse: Transport for London (TfL) – Open data to improve the transportation system

Publishing entity

TRANSPORT FOR LONDON

Use-case overview: TfL – Open data to improve transportation system

Description: TfL Open Data is a program that provides access to a wide range of data related to the transportation system in London. This data is made available to the public for free and is intended to be used to develop new products, services, and applications that can help improve the transportation system in London.

Existing users

App Developers, Researchers, Businesses and Civic Tech organizations

There are ~13,000 registered developers on TfL

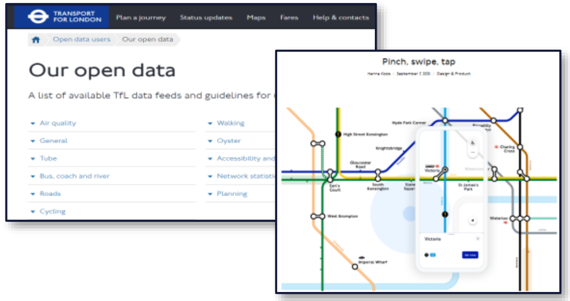
Used datasets:

All open datasets used

Examples:

London underground data for passenger volume and movement

Walking times between adjacent stations



Our open data

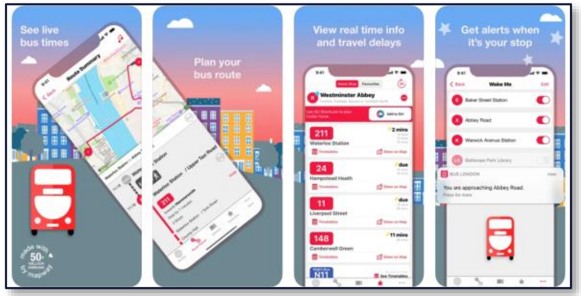
A list of available TfL data feeds and guidelines for

- Air quality
- General
- Tube
- Bus, coach and river
- Roads
- Cycling
- Walking
- Oyster
- Accessibility and
- Network statisti
- Planning

Pinch, swipe, tap

Click to access dataset

Apps developed using TfL Open Data


See live bus times
Plan your bus route
View real time info and travel delays
Get alerts when it's your stop

Key outcomes

Societal impact:

- Efficient transportation system for citizens
- Greater transparency in transportation management
- Better city planning by mapping citizen movement

Economic impact:

- Opportunities for developers and businesses
- Efficient allocation of resources by analyzing areas of immediate improvement

3. Informed decision making, increasing potential value generation

The case of eHealth – Ireland

Overview

- In 2017, Ireland launched eHealth, a platform that brings together **open data** from the **Irish Health Sector**
- The **platform uses**, amongst others, **open data** from the **Department of Health** and from the National Healthlink Project.

  This includes **data on:**
  - Available health services
  - Statistics on hospital cases
  - National waiting lists
  - Key trends on new digital initiatives
  - Prices for medical treatments

Benefits

eHealth uses this open data to **facilitate transparency** in the healthcare sector and to **provide citizens**, **care providers**, and **researchers** with the **information they need to make better decisions**, **spur for new innovations**, and **identify efficiency opportunities**

4. Eased ability to conduct research

Advisory services for the built environment sector through research, testing, and training practices

Overview

- The Building Research Establishment, **provides advisory services through research**, testing, and certifications **for the built environment sector**
- One of the BRE services, BREEAM, or the Building Research Establishment Environmental Assessment Method, **analyzes buildings and projects to make cost effective** and **regulation certified decisions through open data**
- Services include access lists of BREEAM **assessors and assessments**, maps of BREEAM certifications, **data visualizations on projects**
- Sources of **government data** include for example the Department for Communities and Local Government; BRE verifies the **total number of Code for Sustainable Homes certificates**

Benefits

Provides **cost effective and regulation certified resources** for building construction while **reducing the technical and financial burdens** of manually conducting the research

> Government entities should not only publish data but also strive to create applications and services on top of open data to ease access and increase user friendliness

Thank you

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority