Open Data Capacity Development
Module 1: Open Data Across the NSS
Training Syllabus

This training module was developed under the guidance of the United Nations Statistical Division as part of the Data4Now Initiative.

Introduction

This syllabus is intended to guide trainers who are presenting a course on open data for staff in NSOs and other agencies of the national statistical system responsible for documenting and disseminating open data and databases of indicators. It may also be of interest to anyone who wishes to know about the implementation of open data principles.

The training module is part of a larger program on the development of open data capacity in official statistics agencies. It is accompanied by a PowerPoint presentation that can be used for group presentations or for individual learning.

What do I need to know before using this module?

This module provides a broad survey of the principles of open data as applied to the work of national statistical agencies. It does not require prior familiarity with open data or the operation of a statistical agency.

Learning objectives

The expected learning outcomes for this module include:

- Become familiar with the concepts and principles of open data.
- Learn how the principles of open data apply throughout the data value chain.
- Learn the benefits of open data for data producers and users.
- Recognize the coordination challenges and solutions across the national statistical system (NSS) to achieve open data.
- Assess where to start to open up data across the NSS.

Note to trainers: Depending on the pace of the trainer and trainees, it is expected that this module can be delivered in 1-1/2 to 2 hours.

Table of Contents

1. What is open data?

This section introduces the basic definition of open data in the context of official statistics and in the next section describes the benefits of open data to national statistical system (NSS) with examples from countries that have successfully implemented open data. In the following section we will look at each of the open data principles more closely.

1.1 Open data definition

The core concept of open data is defined by the Open Knowledge Foundation as

> "Open data and content can be freely used, modified, and shared by anyone for any purpose."

This means that open data can be acquired for free and without unnecessary hurdles, can be shared by users with other parties, and the content can be modified, for example by changing images, or data series can be deflated, converted to per capita measures, or rescaled in some other way.

The concept of open data has been codified in the Open Definition, which specifies one legal and three technical criteria of openness:

1. Open License (Legal)
2. Open Access (Technical)
3. Machine Readability (Technical)
4. Non-proprietary (Technical)

These criteria will be discussed further in Section 4. Open data principles implementation

1.2 Limitations on open data

Not all data should be made openly available. Data that may reveal personal or proprietary information or that pose security concerns must be more carefully controlled—less open—than anonymized data and aggregate indicators. Data, as shown in **Figure 1** are not open or closed except in extreme circumstances; rather they exist on a spectrum of openness. It is the responsibility of the data steward to determine whether data can be licensed as fully open data.

Figure 1: What types of government data should be open?
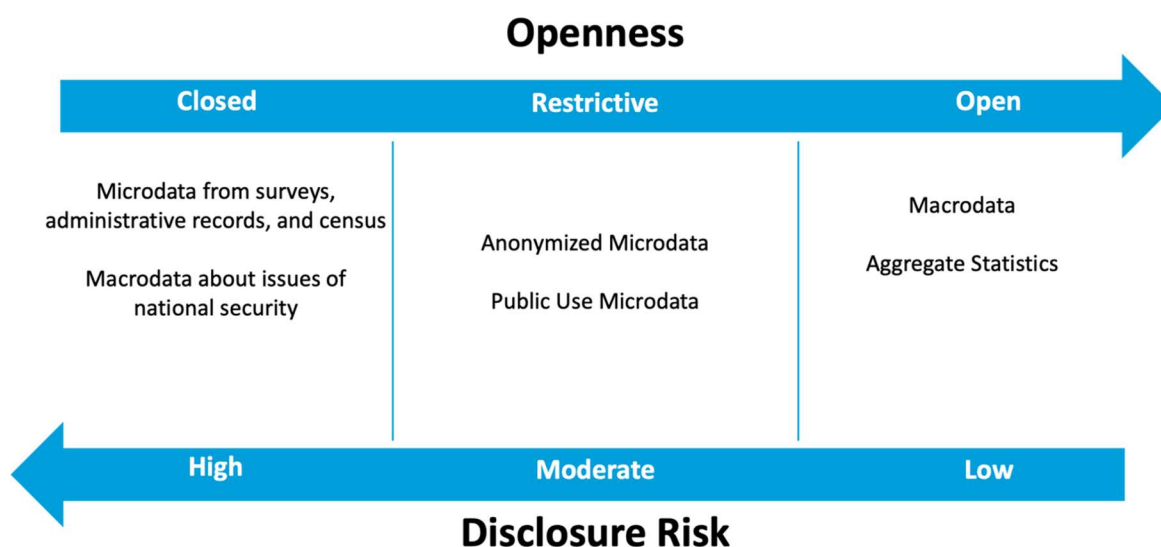
2. Implementing open data involves balancing benefits and challenges

Benefits of open data

- Increasing the use and reuse of data
  Open data initiatives encourage wider access to datasets, fostering innovation and driving economic growth through the development of new applications and insights.
- Improving the value of data
  Implementing open data principles increases the value of official statistics by enhancing transparency and accessibility, enabling more comprehensive analysis and informed decision-making. Open data also unlocks greater value by removing inefficiencies in disseminating data.
- Improving the impact of data
  Open data amplifies the potential impact of data by allowing more users to access data and therefore broadening the base of users that can process and absorb the conclusions from the data. Furthermore, greater impact can also come from open data allowing innovators to link data and therefore facilitate the development of new applications and services based on open data.

- Making the national statistical system more user-centric
Adopting open data principles also makes the entire statistical system more user-centric, creating a virtuous cycle of user engagement with data dissemination efforts and potentially other areas such as improved data collection and user feedback.
- Raising the profile and capacity of the NSO
By promoting transparency and accountability, open data initiatives elevate the status of national statistical offices, enhancing their credibility and relevance to policymakers and the general public.
- Monitoring development progress
Open data enable real-time monitoring of development indicators, facilitating evidence-based decision-making and enabling governments to track progress toward local, national, and global development goals.

Challenges for open data

- Finding resources and building capacity
Implementing open data initiatives can require investment in technical resources and human capacity building to ensure data quality, standardization, and compliance with open data principles.
- Data sharing and interoperability
Ensuring seamless data sharing and interoperability across different agencies and jurisdictions is a continual challenge, requiring the development of common standards and protocols and coordination mechanisms for sustained engagement across teams.
- Privacy and security
Safeguarding privacy and ensuring data security are paramount concerns in open data initiatives, requiring robust data protection measures and policies to mitigate risks associated with unauthorized access or misuse of sensitive information.

3. The Data Value Chain

Producing open data requires an open process. Therefore, open data principles apply across the data value chain. Open data encourages feedback to flow through the entire data value chain, improving operations at every level.

Figure 2: The data value chain
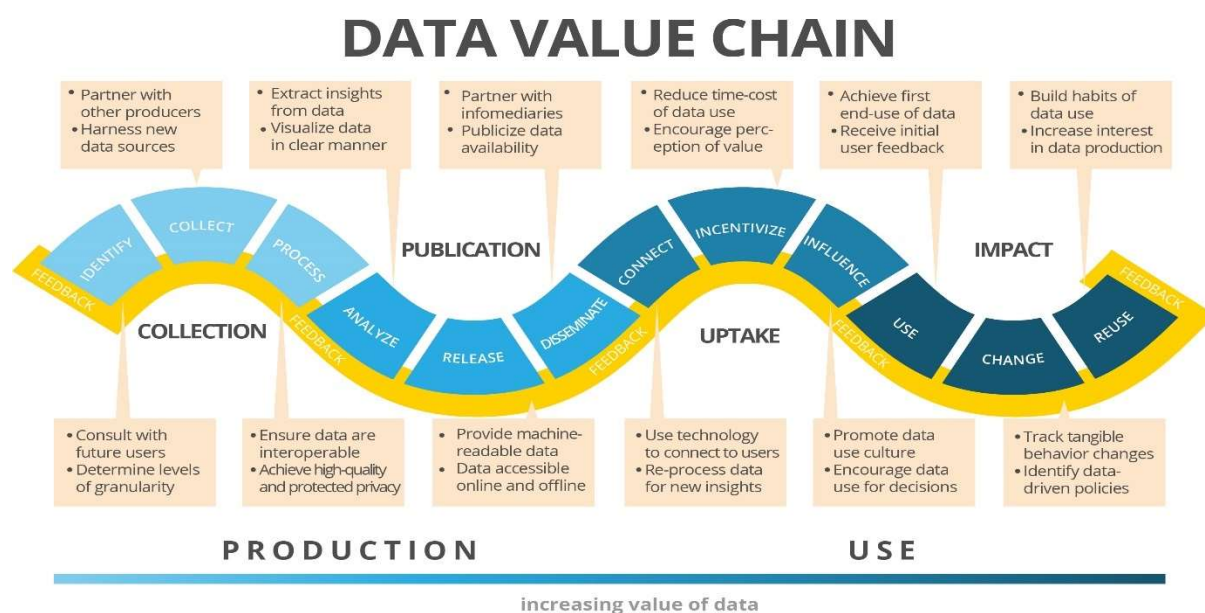
From the beginning of the data value chain, open data require transparent and inclusive data collection processes, fostering trust among stakeholders and enhancing data quality. As data progresses through processing and analysis stages, adherence to principles of openness promotes interoperability, standardization, and accessibility, enabling seamless integration of diverse datasets and facilitating meaningful insights. In the final stages, open data principles require the unrestricted dissemination of information, empowering users to access, reuse, and redistribute data freely, thereby maximizing its societal and economic value and leading to greater impact. Embracing openness throughout the data value chain not only enhances transparency and accountability but also stimulates innovation and drives evidence-based decision-making, ultimately advancing the collective pursuit of knowledge and progress.

4. Open data principles implementation

This section examines how open data principles are implemented and how they are assessed using the Open Definition as applied by the Open Data Inventory (ODIN).

4.1 Adopting Open Data Principles

Open by Default

The open by default principle serves as the foundation for a set of policies that make a government's or an organization's data publicly available and in accordance with open data guidelines, with only a limited number of specific exceptions (for reasons of security, for example, or privacy protection). Under the principle, it is recognized that government data produced with public resources are valuable and have many different users and uses and should therefore be expected to be available to the public (that is, by default).

The UN Statistical Commission at its 53rd session proposed a general framework for implementing an open by default framework within national statistical systems:

1. Establish cross-governmental support for open by default policies.
2. Adopt a public domain dedication or open data license.
3. Show transparency regarding data sets that cannot be released publicly
4. Establish a transparent data request process
5. Establish clear procedures for microdata access and publication
6. Publish a data publication timetable
7. Build staff capacity to produce open data
8. Build external users' capacity to use and understand open data

This list requires the investigation and coordination with national legal frameworks, for example copyright laws, governments data licenses, national statistical laws, among others. In addition, statistics producers may have concerns that data may be misused, that their reputation may be hurt by open data, that there may be questions of individual privacy, and they may fear the loss of revenue. Each of these concerns can be mitigated by clear communication about the implementation of the open by default approach within the specific context of each statistical office.

Incorporating a user-centric focus in open data

User-centered design is about engaging users at all stages of platform development and data dissemination. When user-centered design is done well, it can increase data use and

deliver the right data in the right formats to a wider group of stakeholders, starting a virtuous cycle of data use that supports the open data movement at large. As decision-makers find the datasets they need through well-designed and open data platforms, they will make better decisions, leading to better outcomes and more support and funding for creating and opening of data.

The principles of open data were created in response to user needs and are a foundation of user-centric design. As data producers continue to learn from and respond to users, the standards for open data will evolve and be operationalized through user-centric design. To better understand their users, data producers should:

1. Develop a user-engagement strategy and streamline user feedback;
2. Implement web analytics to identify users' activities;
3. Conduct surveys, focus groups, and meetings to gather users' needs.

These actions are best practices to increase the use of data. If data producers listen to users, they will design better products and provide not just open data, but useful open data, further improving the value of their data.

National reporting platforms for open data

Most users access open data through national reporting platforms (NRPs), which may be simple websites or sophisticated data portals. Well-designed and well-managed NRPs are governed by a set of general principles that support access to and use of open data. These principles were set forth in the report on [Principles of SDG Indicator Reporting and Dissemination Platforms and Guidelines for Their Application](#)

Principle 1: Clear institutional arrangements and management

Data producers should provide documentation of the laws, regulations, mandates, and standard operating procedures that govern the development, implementation and maintenance of the national reporting platform. Good documentation increases transparency and reduces confusion over the responsibilities of different actors in the NSS and their roles in managing the NRP, leading to increased efficiency and a better managed NRP and greater confidence of users.

Principle 2: Fitness for purpose

A national reporting platform should be established in cooperation and consultation with all stakeholders, in compliance with the Fundamental Principles of Official Statistics, to address their needs for subnational, national, regional, and global reporting. An NRP that is

aligned with the goals and needs of stakeholders will be fit for purpose. (See Module 4 *Storytelling with Open Data* for further discussion on how storytelling can be used to meet the needs of stakeholders.)

Principle 3: Sustainability

The purpose, scope, features, development, and implementation of a national reporting platform should take into account the resources and capacities required over the medium and long term. National statistical offices (or whomever is responsible for the platform) and partner organizations should consider specific objectives of the country in ensuring the sustainability and support of the national reporting platform.

Principle 4: Interoperability and statistical standards

The NRP should adhere to statistical standards in the development of its products by modeling the data production process and adopting common data structure definitions and code lists. The use of common standard supports interoperability of data systems benefits the NSS through increased efficiency in data flows within the national statistical system and increased usability for users.

The purpose of the NRP principles is to guide NSOs in incorporating best practices for usability into their platforms. NSOs typically assume leadership in the national statistical system, where they are a critical actor in the implementation of standards. Depending on the structure of the national reporting platform, the role of the NSO may vary, although the NSO is typically the primary organization in managing the platform, albeit with careful coordination with key producers of statistics within the national statistical system. The principles and guidelines are also integral to the mission of national statistical offices worldwide – to publish data that are available to all users. (See Module 3 *Open Data Portals* for a further discussion of how data portals support open data.)

4.2 Measuring data openness?

Open data conform to the standards of the Open Definition. The Open Data Inventory (ODIN) produced by Open Data Watch is a biennial evaluation of the coverage and openness of data provided on the websites maintained by national statistical offices (NSOs) and any official government website that is accessible from the NSO site. The overall ODIN score is an indicator of how complete and open an NSO's data offerings are.

ODIN looks for data in 22 categories of official statistics represented by 65 indicators. It evaluates these data for availability (coverage) and openness. Here we focus on openness criteria.

How does ODIN define and measure openness?

ODIN implements the four criteria of the Open Definition and complements them by adding the availability of metadata as a fifth element.

Machine-readability

Are the data published in a machine-readable format? Machine-readable formats allow users to easily process data using a computer. When data are made available in formats that are not machine readable, users cannot easily access and modify the data, which severely restricts the scope of the data's use. XLS, XLSX, CSV, Stata, SAS, SPSS, JSON, CDF, RDF, XML, and TXT files are classified as machine readable. PDFs, HTML, DOC, DOCX, PPT, PPTX are common formats that are not machine readable.

Non-proprietary

Are the data published in a non-proprietary format? Nonproprietary formats are important because they allow users to access data without requiring the use of costly and exclusive software that may prevent some users from accessing the data. PDF, HTML, XLSX, DOCX, CSV, JSON, XML, and TXT files are classified as nonproprietary. XLS, Stata, SAS, SPSS, DOC and PPT files are common proprietary files.

Open Access

Can the data be downloaded in bulk or with custom-selection? ODIN looks for the presence of three different download options: bulk download, API, and user-select download options.

Bulk downloads are a key component of the Open Definition, which requires data to be "provided as a whole...and downloadable via the internet." In ODIN, a bulk download is defined as: The ability to download all data for a particular indicator at the national level (all years and disaggregations) in one file, or multiple files that can be downloaded simultaneously.

API stands for Application Programming Interface. An API is a tool that allows external applications, such as analytical software or website plugins to directly access data. Ideally, an API option should be clearly displayed on the website with directions on how to use it.

User-select download options are defined as a process by which users can select an indicator and at least one other dimension to create a custom table. These dimensions could include time periods, geographic disaggregations, or other types of disaggregations.

Metadata

Are the data published with metadata? Although not explicitly mentioned in the Open Definition, metadata are essential for the open use of data. Metadata will also improve interoperability to link datasets across the data holdings of the NSS.

Many metadata schemas exist (See Module 2 *Open Metadata* for more information) but at a minimum, each dataset should include:

- Definition of the indicator or definition of key terms used in the indicator description (as applicable), or how the indicator was calculated.
- Publication date or date the dataset was last updated.
- Name of data source or the agency that collected the data.

Open License

Open data license are the bedrock of an open by default approach. "Terms of use" and "data license" are both used to describe the policies of data producers concerning how their data can be used. Adopting and publishing an open data license encourages greater use of data by clearly identifying how people can use data. This alleviates concerns about economic or legal sanctions for unapproved use.

A license or terms of use is considered open if it explicitly allows the use and reuse of data for any purpose, including commercial and noncommercial use. The only obligations that can be required of the user are attribution of the data to the original source or requiring users to relicense the data under the same or similar open license (called a "share-alike" requirement). Other requests can be made of the user but must be suggestions and not enforceable under the license or terms of use.

Requirements not considered acceptable under the Open Definition are requiring permission before any kind of use or general provisions forbidding misleading or inaccurate use of data or other attempts to penalize broad categories of misuse that can be arbitrarily applied and have a chilling effect on all data users. Furthermore, the data license should not be used to penalize illegal activities already forbidden under law.

The data license should also not require users to publish metadata with any dataset they reuse, beyond information regarding any changes the user made to the dataset or a link to the original dataset.

Statistical agencies can adopt a standard license such as those provided by Creative Commons (CC). The most commonly used is the CC BY 4.0, which requires attribution, and the CC0 or public domain license. Some governments have also created their own

open data licenses following the guidelines in the Open Definition. The data license may be included as part of a broader policy on access to government information, but the data license should always be readily available to data users.

Country Examples of Open Licenses/Terms of Use

St. Lucia: This license format is used by many countries in the Caribbean and is easy to follow. It addresses many legal issues and provides examples of how users should cite their data.

Australia: ABS Stat uses a Creative Commons BY 4.0 license, clearly lists the exceptions, and also links to the full legal code of the license, as well as the country's Copyright Law.

Korea: They provide a clear description of the various access levels of data in their portal, as well as contact information in case users have questions about the public use of certain datasets.

5. How do countries perform?

For the ODIN 2022/2023 round of assessments (**Figure**), the highest scoring openness element was non-proprietary data, meaning that many countries find it easy to publish data in at least one format that is free to use. This element was followed by machine-readability and metadata availability, the latter having made large improvements since 2020. Download options were scored around 40 out of 100, indicating that many users receive little assistance in understanding their data.

Figure 3: Median openness element scores from Open Data Inventory (2020-2022)
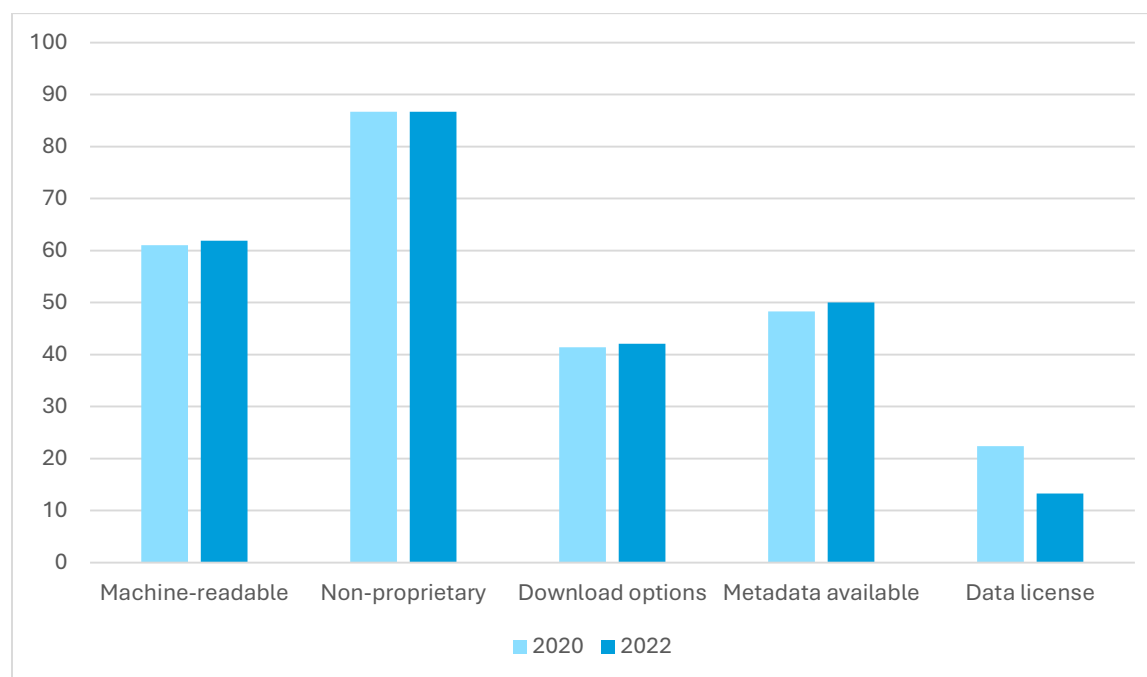
We take special note of the importance of data licenses because they scored worst among all openness elements in ODIN and scores decreased further in 2022. That year, 78 countries scored 0 for their data license across all categories, indicating that none of their data were released with an open data license. Most of these countries have no data license at all, suggesting that they lack awareness of the importance of data licenses. But this may be changing. In March 2022, the UN Statistical Commission endorsed the report of the Open Data Working Group that recommended the adoption by national statistical offices of the principle of "open data by default" of which open data licensing is an essential component.

Many countries have already made changes. Several countries that improved their score adopted entirely new terms of use or data licenses for their websites—for example, Sierra Leone and Jordan (See ODIN 2022 report). Previously, Sierra Leone's terms of use were vague and did not specify how people could use their data, and Jordan's did not allow commercial use of their data. Both countries adopted new licenses without these limitations. Jordan adopted a new bespoke license now used on its NSO website and many of its data portals. Sierra Leone took a different approach, adopting an open data template used by many other countries and customized it to meet their purposes.

Other countries simply edited their existing licenses to align with the Open Definition. Malaysia is an example. Instead of adopting an entirely new license, their existing license was edited to remove a statement that required users to ensure that the data they use are republished "accurately." Vague requirements such as this are found in many countries' data licenses. Because they are not explicit about what is prohibited (or in this case, what is considered "accurate"), they create a legal grey area that could potentially be used to restrict the use of data or punish data users. These requirements should not be found in open data licenses.

6. Coordination across the NSS to achieve open data

Countries find coordination a challenge if the institutional arrangements of the statistical system, whether by statistical laws or other organizational mandates make coordination difficult across the offices of the NSS. Statistical laws or regulations can inhibit coordination in statistical systems by creating fragmented data collection and reporting processes, imposing rigid bureaucratic structures, and restricting data sharing across different government agencies. For instance, laws that mandate strict confidentiality may prevent the sharing of vital information necessary for a cohesive statistical system. Additionally, the absence of a central coordinating body or clear guidelines can lead to overlapping responsibilities, inefficient use of resources, and difficulty organizing collective efforts around open data.

To address these challenges, statistical laws and regulations should encourage inter-agency collaboration and establish frameworks for data sharing while maintaining data privacy and security. Examples of laws that mandate coordination include the Philippines' "Philippine Statistical Act of 2013," which created the Philippine Statistics Authority to centralize data collection and dissemination, and the "National Statistical System Act" of Kenya, which establishes the Kenya National Bureau of Statistics as the central coordinating agency for statistical activities. These laws provide a legal basis for improved coordination and collaboration across various statistical agencies, leading to more efficient and effective statistical systems.

Increasingly, countries are finding the role of a data steward to be appropriate to govern the intricacies of a national statistical system. The responsibilities of stewardship can be concentrated in one person or one office or can be shared. Within the context of open government data initiatives, data stewards, by implementing standard protocols and best practices, can streamline data collection processes and minimize errors or discrepancies

in publicly available datasets. Furthermore, effective data governance frameworks promote collaboration and knowledge-sharing among government agencies, civil society organizations, and citizens, fostering a culture of transparency and participatory governance. (See Module 5 *Citizen-generated data and open data* for a discussion on how citizen-generated is being integrated into official statistics through governance processes.)

Data exchange frameworks facilitate seamless sharing of information across government departments and agencies, enabling policymakers and researchers to access diverse datasets for analysis and decision-making. By promoting collaboration and interoperability among stakeholders, these mechanisms contribute to the harmonization of data standards and methodologies, thereby improving the quality and reliability of statistical information. Moreover, open data initiatives supported by robust data exchange mechanisms promote accountability and public trust by allowing citizens to scrutinize government activities and monitor progress towards development goals.

7. How to open up data across the NSS

1. Adopt a policy of open by default with a narrow and explicit list of exceptions

Start by establishing a national policy that mandates all data produced by the government be open to the public by default. This policy can specify clear and narrow exceptions to this rule, such as data that, if released, could compromise national security, violate personal privacy, or infringe on intellectual property rights. Such a policy ensures a culture of transparency and sets a clear expectation that government data should be freely accessible unless there is a compelling reason otherwise.

2. Ensure all data are published in nonproprietary and machine-readable formats

It is crucial to publish data in formats that are open, nonproprietary, and machine-readable, such as CSV, JSON, or XML. This enables a wide range of users, including researchers, developers, and policymakers, to easily access, analyze, and repurpose the data. Avoiding proprietary formats ensures that the data can be freely used without requiring specific software, thus maximizing its accessibility and usability.

3. Publish metadata consistently for all datasets and ensure metadata standards are followed

Consistent publication of metadata—information that describes the data, such as its source, methodology, and date of collection—is essential for users to understand and

trust the data. Implementing and adhering to established metadata standards ensures that all datasets are accompanied by comprehensive, standardized descriptions, facilitating easier discovery, comparison, and integration of datasets across different platforms and sectors.

4. **Adopt an open data license**

Implement an open data license, such as the Creative Commons Attribution License (CC BY), to clarify how the data can be used. An open data license provides legal certainty for users by explicitly stating the terms under which data can be accessed, shared, and reused. This encourages broader use of the data by removing legal ambiguities and fostering a culture of openness and innovation.

8. Key Takeaways

- According to the Open Knowledge Foundation's Open Definition, "Open data and content can be freely used, modified, and shared by anyone for any purpose."
- Open Data have several benefits, such as increasing the use and reuse of data and thereby improving the value and impact of data, together with making the national statistical system more user-centric, raising the profile of the NSO, and helping to monitor development progress.
- Open data principles apply across the data value chain. Open data encourages feedback to flow through the entire data value chain, improving operations at every level.
- The principles of open data implementation are Open by Default, User centricity, and open national reporting platforms.
- Adopting an Open by Default framework can involve many steps but most fundamentally requires adopting an open license.
- The Open Data Inventory (ODIN) assesses the machine-readability, non-proprietary status, download options, availability of metadata, and terms of use of a diverse set of development data.
- Countries have improved on ODIN openness components over time but data licenses and open data licenses in particular are lacking for many countries.
- Improving open data is not just a technical question but also requires overcoming fundamental governance and coordination hurdles that impede sharing data.