Open Data Capacity Development
Module 3: Open Data Portals
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

- Why data portals are an important part of an open data program
- What are the United Nations Statistics Division's Principals and Guidelines for data platforms?
- How existing data portals be improved and made more open
- Some examples of open-source data portals and their use

Table of Contents

1. Data portals for open data

The principles of open data require data to be accessible to all users in convenient and modifiable forms. This means open data should be easily discoverable with sufficient documentation to inform users of the content and quality of the dataset. Data providers can increase the usability of data by providing data through data portals with a well-designed user interface. Data portals increase the transparency of the national statistical system and can increase trust in data. They can also benefit the NSO by increasing the efficiency of data management by improving data exchange and through integration with underlying databases.

> Today, data portals form a critical part of the infrastructure of official statistics: they are the principal link connecting data users and data producers. They perform a range of functions as platforms for data publication, data discovery, monitoring of public policies, and engagement with data users and citizens, among others. Data portals are key enablers of data access and use. They serve a vital function in the overall virtuous data cycle by which good data encourages greater use leading to improved policies and outcomes, and which, in turn, increases demand for data. (*Data Dissemination in the Digital Age*, PARIS21 and Open Data Watch, 2021.)

What is a data portal?

Data portals exist in many forms for handling different types of data, but they all share the common purpose of making complex data sets more easily accessible. The United Nations Handbook of Statistical Organization describes a data portal as "...a web-based, interactive data and metadata platform with databases modelled for specific data types and domains such as microdata, macrodata or geospatial data." (4th Edition of the Handbook of Statistical Organization, United Nations Statistical Division)

The *Open Data Handbook* of the Open Knowledge Foundation describes some of the important features of a data portal: "The aim of a data portal is to provide a data catalogue, making data not only available but discoverable for data users, while offering a convenient publishing workflow for publishing organizations. Typical features are web interfaces for publishing and for searching and browsing the catalogue, machine interfaces (APIs) to enable automatic publishing from other systems, and data preview and visualization."

Overcoming obstacles to data use

Although the web has become the standard means of publishing and sharing information, many countries continue to rely on less sophisticated means of disseminating data, such

as PDF copies of statistical year books, ad hoc links to data tables, or outdated dissemination platforms that lack full functionality. Data published through these methods are less findable and less usable.

Other obstacles to effective, open data dissemination are websites or portals that do not support local languages, portals that experience lengthy downtime, and the proliferation of many portals with differing interfaces and content or, in some cases, similar but unsynchronized content that leave potential users uncertain about the provenance and reliability of the data.

2. Principles and guidelines for data portals

In 2018, the United Nations Statistics Division published 4 principles and 12 guidelines for the implementation of open and effective data portals. Although directed at the development of platforms for reporting SDG indicators, the principles are applicable to general data portals. In this section we examine the principles and guidelines and their application to the development of national data portals.

(In the following sections, the text in *italics* is taken from the Principles of SDG Indicator Reporting and Dissemination Platforms and Guidelines for Their Application (Annex 1 to *Report of the Conference on National Reporting and Dissemination Platforms*, New York, 22-24 January 2018)

Principles

Principle 1. Clear institutional arrangements and management

The responsibility for the development, implementation and maintenance of a national reporting and dissemination platform, including mechanisms of coordination and cooperation within the national statistical system should be clearly established.

When the platform is established by the National Statistical Office, a strong mandate for coordination of statistical activities and data exchange agreements helps to ensure a timely and efficient compilation of data from different sectors and proper management of the portal.

Principle 2. Fitness for purpose

The exact purpose, scope, and features [of the platform] should be established in cooperation and consultation with all stakeholders and users. Platforms should comply with the Fundamental Principles of Official Statistics and should address the priority needs and requirements of subnational, national, regional, and global monitoring and reporting.

Fitness for purpose recognizes that different data users have different needs. Some may require nothing more than access to the data and metadata. Others require visualizations and interpretations of the data. Filtering mechanisms make it possible for users to select the data they need.

Principle 3. Sustainability

The decision on the purpose, scope, features, development, and implementation [of the platform] needs to take into account the availability of internal and external resources and capacities over the medium and long term to ensure sustainability of the platform.

Too often, well-intentioned programs end with orphan platforms, unsupported and out-of-date.

Principle 4. Interoperability and statistical standards

The components [of the platform] should follow international and national statistical standards and best practices, to facilitate the integration, customization and further development of components and different solutions. In particular, standards should be supported that aim to facilitate data harmonization and exchange across different stages of the statistical production and across institutions, including common data structure definitions and code lists.

All data portals are mechanisms for sharing data and should be designed to facilitate exchange and interoperability of data.

Guidelines

The document also provides guidance on the implementation of national data platforms.

National ownership

National statistical offices should have the ability to maintain, adapt, transform, and customize their data portals to address their own needs and the needs of their users.

In many cases, countries may rely on data portals that have been developed and managed by donors or other parties. Such platforms should, nevertheless, allow for local customization. National authorities should retain control over the content of the portal.

Collaboration

Data portals should be designed, developed, improved, and maintained using a collaborative approach that leverages learning between various stakeholders of the national statistical system as well as technology developers, donors, policy makers,

subject-matter experts, business partners, advocacy groups, and both institutional and grassroots users.

Implementing a data portal is not simply an IT issue. Collaboration in the design of a data portal ensures that the needs of those responsible for producing and managing data and the interests of those who will use the data are considered in the design and operation of the portal.

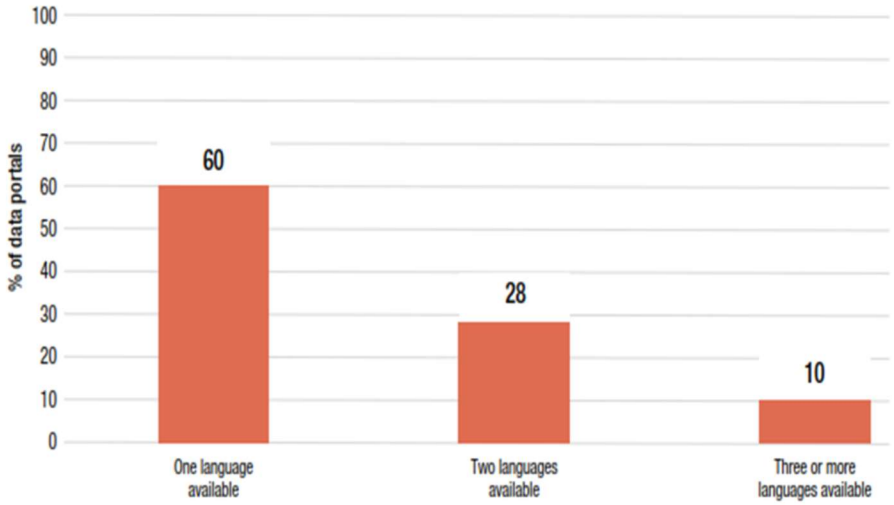
Another opportunity for collaboration is through open-source software developed by others. Open-source portals that can be used by other agencies within the NSS or shared with other countries are a more efficient and sustainable way of developing and maintaining a data portal.

Multilingualism and accessibility

To leave no one behind, ensure national ownership, and promote the use and impact of data for policy and decision making at the local level, data portals should support national languages and implement national and international best practices in terms of accessibility to persons with disabilities, as well as full access across the range of browsers and devices, including mobile devices.

The 2021 study of data portals by ODW and PARIS21 found that 60 percent of portals in IDA-eligible countries were available only in one language, usually English. Languages that were available were usually major international languages, not national or local languages. The study also found problems with the accessibility of portals for people who are visually impaired.

Figure 1 Data portal language options



Source: Data Dissemination in the Digital Age, 2021

User-centered design

Data portals should be designed for and with users (including both operational and end users, such as data consumers or NSO officers), and project owners should engage them in all phases of development. … Regular collection and analysis of usage data and online user feedback should further assist in providing guidance for future modifications and enhancements.

Encouraging, acknowledging, and making use of feedback demonstrates a commitment to open design and open data. This can be done through online responses from the portal and by web analytics that provide quantitative information on its use.

Data communication

Data portals … should support multiple ways to explore, represent and communicate data on statistical indicators, and address the needs and priorities of diverse groups of users … This includes innovative data visualization and data storytelling capabilities.

There are many ways to communicate data. Giving users options for visualizing data through charts or maps will increase understanding and encourage data use. To further increase use, portals should support tools for sharing data and visualizations on social media.

Data disaggregation

Data portals should support improved access to, and use of, disaggregated data to focus on all segments of the population, including the most vulnerable.

Portals should provide options for filtering data by administrative unit and by all relevant characteristics, such as sex, age, wealth or income group, disability, ethnicity, or migrant status.

Modularity and extensibility

Data portals should be modular, composed of modules (sub-systems) and components that interoperate to service the different phases of the data life … The data that these modules and components consume as inputs and produce as outputs should, as much as possible, be based on open standards and protocols such as Statistical Data and Metadata eXchange (SDMX). … The systems should support extensibility through the addition of modules or components, upstream or downstream.

Modular design requires a comprehensive data model that describes the flow of data from one processing stage to the next. SDMX supports the exchange of data and metadata in a standard format to increase interoperability with other systems. Internally it encourages documentation of the data production process, increasing the quality and reliability of data.

Standardized interfaces

Data portals should provide standardized APIs in accordance with the industry's best practices, such as the OpenAPI Specification. This facilitates creating and sharing data across global, regional, national, and subnational data communities.

APIs (application programming interface) are a feature used for sharing data with other programs. To encourage their use, they should follow standard practices and be well documented on the data portal website.

Scalability

A national reporting and dissemination platform should have an architecture that enables a statistical office to start with a limited scale implementation and iteratively progress towards a full-scale system. Tasks such as adding indicators or breakdowns should be performed directly by an operational user at the statistical office and should not require additional software development.

No system is built in one day nor should a data portal be built for only one purpose. By adopting a collaborative and user-centered approach and a modular design, data portals

can be built and managed to accommodate a variety of data types and users as their needs evolve.

Metadata

Data portals should support statistical metadata at the appropriate level of granularity. This includes structural metadata such as codes and their descriptions; reference metadata such as methodology and quality aspects of published indicators; and other relevant information such as the date of the last update.

Adequate metadata are a key feature of open data systems. They define each indicator, describe how data are transformed through the statistical production process, and make the data findable. Without metadata, the data hosted on a data portal may be unusable. A sophisticated data portal makes metadata accessible with the data, but simpler systems – a spreadsheet or a PDF file – can still provide users with the information they need.

Open data

Data portals should be consistent with open data best practices, summarized as "Open data and content can be freely used, modified, and shared by anyone for any purpose." National reporting and dissemination platforms should include and follow a data license consistent with the open data principles, such as Creative Commons Attribution (CC-BY 4.0) or the Open Database License. Published datasets should be clearly attributed to the originating organization

Lacking a data license, users may be reluctant to use data because of concerns about the legal consequences of unapproved use. The Open Data Inventory (ODIN) assesses the openness of official data websites and portals. These sites are often marked down because they do not offer good options for downloading data, lack sufficient metadata, and do not include an open data license

Linked data

Responding to the increasing demand for open data to leverage global and national investments in data for evidence-based policy and decision making, data portals should comply with a minimum of Level 3 of the 5-star Linked Open Data Principles (Berners-Lee, 2012)

Figure 2 Five-star open data principles

★	Make your stuff available on the web (whatever format)
★★	Make it available as structured data (e.g., excel instead of image scan of a table)
★★★	Non-proprietary format (e.g., CSV instead of Excel)
★★★★	Use URLs to identify things, so that people can point at your stuff
★★★★★	Link your data to other people's data to provide context

Source: https://inkdroid.org/2010/06/04/the-5-stars-of-open-linked-data/

The first three levels of the 5-star open data principles would be met by any data portal satisfying the standard requirements for open data: provide all data in machine-readable and structured, non-proprietary formats. At the more advanced levels, Berners-Lee calls for using the resource description framework (RDF) to make data available through the semantic web. The ODW-PARIS21 study recommends, "As there are many other easier to implement changes needed to adhere to the principles and guidelines, these should be prioritized before incorporating linked data features."

3. Improving the performance of open data portals

None of the 78 data portals examined by ODW and PARIS21 fully implemented the UNSD principles, and almost a third of IDA-eligible countries did not have a data portal at all. For countries with portals, the study found that many could be improved through simple and inexpensive steps, such as adding an open data license, providing more metadata, and offering a bulk download option.

The study made five recommendations for building or improving national data portals.

1. **Prioritize a back-to-basics approach** by first implementing features that will provide a high return on investment. For example, improving metadata availability and multilingualism should be implemented before more advanced features such as APIs and linked data.

2. **Adopt user-centric design** to account for the needs of end users. User feedback combined with quantitative methods such as Google Analytics and qualitative methods such as focus groups and interviews can all be used to produce a picture of user needs along with success stories to encourage other users.

3. **Advance national ownership** of data portals. Even when external partners aid the design and implementation of a portal, NSOs should be fully engaged in the choice of a portal and the dissemination of national data. Development partners should provide training and engage in regular consultations to bring together country actors to assess priorities.

4. **Improve upstream data management** practices for a sustainable data dissemination infrastructure. The long-term sustainability of a data portal depends on robust data management practices. For example, data modeling and implementation of data standards should incentivize a shift from paper to digital methods, increasing data quality and organizational efficiency.

5. **Streamline data dissemination processes** to reduce the reporting and management burden for maintaining data portals. Integrated, yet modular, "all-in-one" data portals and streamlining data dissemination processes could reduce the reporting and management burden on NSOs. Where there is more than one portal, they should be interoperable and integrated as much as possible to minimize the burden on data providers and ensure that users do not have to search many portals to find the data they need.

4. Open-source data portals

Open-source software is the programming equivalent of open data: it is computer software that is made available with a license that grants users permission to use, change, and redistribute the source code to anyone or for any purpose. The ability to examine the source code increases trust in the code and its applications. One of the advantages of using open-source software is that successful software may have a large community of users to improve it or adapt it for other applications.

Open-source software is not necessarily costless. Although the source code may be distributed freely, users may incur expenses installing, customizing, and maintaining it. These services may be provided internally, or it may be more efficient to hire an external provider.

Three open-source data portals have been widely used by statistical offices for publishing macro (indicator) data on the web. PxWeb was developed by Statistics Sweden; complementary tools have been built in collaboration with Statistics Finland. .Stat Suite was developed by the OECD and is now maintained by the Statistical Information System Collaboration Community (SIS-CC), a partnership of 15 international and national statistical organizations. Open SDG

PxWeb

PxWeb is a free web-based table distribution system for statistics producers. It has been installed by 111 national statistical agencies in 46 countries, primarily in Europe, Africa, and Southeast Asia. Data published through PxWeb are stored as Px files. Px files can be processed using PxEdit, a program developed by Statistics Finland. Data may also be retrieved from a SQL database.

Users are responsible for their own installation, although some support may be obtained through technical assistance from SIDA. Community support is available from a user Google Group. More information and links to the source files can be found on the pages of Statistics Sweden.

In a typical PxWeb installation, indicators and their dimensions are selected from a menu. Data are retrieved in a table. A pivot function permits the table to be displayed in different orientations and basic arithmetic functions can be used to transform the data. The resulting table can be downloaded in a variety of formats including comma or tab delimited files, Excel, and Json. Options are available for showing results as charts. Queries can be saved and an 'API can be used to import the data into another web-based application.

An example of a PxWeb implementation can be seen on Ghana's StatsBank.

.Stat Suite

.Stat Suite is a standard-based, componentized, open-source platform for the production and dissemination of statistical data. It is based on the General Statistical Business Process Model (GSBPM) and the Statistical Data and Metadata eXchange (SDMX) standards. Development of and support for .Stat Suite is provided by the SIS-CC community. The .Stat Academy provides free training courses, however organizations are responsible for their own implementation.

.Stat Suite has three main modules:

1. .Stat Data Explorer is a front-end for the dissemination of databases. It provides functions for searching, visualizing, and sharing data using an SDMX API. It provides multilingual support
2. .Stat Core is the backend supporting SDMX-based data flows through the statistical organization.
3. .Stat Data Lifecycle Manager is a front-end for the management of the data and referential metadata according to the GSBPM processes of Design, Collect, Process, and Disseminate.

For an example of the .Stat Explorer, see Thailand's Statistics Sharing Hub.

Open SDG

Open SDG is a free, open-source platform for managing and publishing data and statistics related to the Sustainable Development Goals. The platform is customizable and can be used to publish national and global indicators. It supports multilingual displays of data in tables, charts, and maps.

The code for setting up an Open SDG platform is available on GitHub. The current release version is 2.20. The development team can be contacted at opensdg@outlook.com. Because Open SDG is an open-source platform, contributions to the code, design, and documentation are welcome.

Open SDG has been used by countries and cities. For example, see Ghana's SDG platform: https://sdgs-ghana.github.io/ or the city of Johannesburg site: https://sdg.joburg.org.za/. In Central Asia, five countries have merged their data on one platform: https://central-asia-sdg-platform.github.io/www/.

5. Summary and key takeaways

- Data portals increase the transparency of the national statistical system and can increase trust in data. They benefit the NSO by increasing the efficiency of data management by improving data flows within the organization and facilitating data exchange.
- The UNSD has articulated four principles for the development of national reporting platforms:
    - Principle 1. Clear institutional arrangements and management
    - Principle 2. Fitness for purpose
    - Principle 3. Sustainability
    - Principle 4. Interoperability and statistical standards
- Data portals should include and follow a data license consistent with the open data principles, such as Creative Commons Attribution (CC-BY 4.0) or the Open Database License.
- The long-term sustainability of a data portal depends on robust data management practices, increasing data quality and organizational efficiency through the modeling of data flows and implementation of data standards throughout the organization.
- Data portals and supporting tools are available as open-source software with a license that grants users permission to use, change, and redistribute the source code to anyone or for any purpose. The ability to examine the source code increases trust in the code and its applications. Active communities of users support further development of the software.