

Module 7

How to create a
data inventory

Checklist

How to create a data inventory

Contents

Purpose of this guide

When to use this guide

Introduction

Why create a data inventory?

Steps to building a data inventory

- Plan the inventory

 - Decide on the dataset's attributes you want to collect

- Populate the inventory

 - Publishing the inventory so others can find it

- Updating the inventory

Further resources

Appendix: Data inventory template

Purpose of this guide

This guide is intended to support grantees to facilitate data being as findable, accessible, interoperable and reusable (FAIR) as possible by creating a clear inventory of data to help effectively locate, manage, use and share datasets within investments funded by the Bill & Melinda Gates Foundation.

It provides a summary of the benefits of cataloging data in a data inventory, an outline of the steps to create a data inventory and recommendations about what to include in an inventory and how to publish.

When to use this guide

Start concept | request proposal | **refine proposal** | create agreement | request approval | obtain signatures | **active**

You can use this guide in a number of ways;

- At the beginning of an investment, to consider and document the types of data a project will need to access, use and share
- Throughout investments as a tool to help locate, manage, use and share datasets
- In conjunction with documenting and mapping the stakeholders and value exchanges across a data ecosystem, to better understand the 'data landscape' of a domain – what data is out there and how much of it could be accessed, used and shared to solve a given challenge.

Introduction

Grant-making institutions like the Bill & Melinda Gates Foundation aim to create the widest possible benefits from the research they fund. The Agricultural Development program supports country-led inclusive agricultural transformation across Sub-Saharan Africa and South Asia. Within this program, the foundation's Digital Farming Services (DFS) portfolio works to support innovation in integrated digital farmer services to accelerate productivity and income growth for smallholder farmers (SHF).

Creating digitally-enabled services to reduce risk and improve farm-level decisions requires sustainable access to data in order to ensure services are made available and shared with the largest audience possible. The DFS portfolio goal of 'at least 50% of SHFs in target geographies using digitally-enabled services to reduce risk

and improve farm-level decisions' by 2030, can only be achieved when grantees, regional implementing partners, and local institutions are able to find, access, use and share relevant data.

¹Dodds, L. & Wells, P. (2019). Issues in Open Data. The State of Open Data: Histories and Horizons. <https://stateofopendata.od4d.net/chapters/issues/data-infrastructure.html>

Data assets, such as datasets, identifiers, and inventories, are part of our **data infrastructure**¹ and as such need sustainable access and adequate governance. Cataloging data can help to effectively locate, manage, use and share datasets to achieve project or organisational goals.

A data inventory is a list of datasets annotated with important information (known as metadata) that can help users understand why data has been collected, what it contains, how it is managed and the ways it will be made available for others to use. It is a useful tool for any organisation or project dealing with multiple types and sources of data.

When published under an open licence, a data inventory can also help people outside of an organisation to find and use the data they need. Using an open standard for metadata facilitates the discoverability and aggregation of datasets from multiple sources. The resources section of this guide contains further information on this.

Creating a data inventory is also an important part of designing a data management plan.

Why create a data inventory?

A data inventory can help provide useful information on the location, quality, technical and legal frameworks that will inform how data is managed, used and shared. They can also be used as a tool to help drive efficiencies and achieve project or organisational objectives. For example, cataloging data in an inventory can help to:

- **Inform decision making and create efficiencies.** Assessing data assets helps to prioritise resources, improve data quality, rationalise technical platforms used to manage, access and share data, and avoid duplication in collecting or purchasing data that is already available.
- **Tackle a specific problem.** An inventory might also be compiled to provide a list of datasets that are useful in [tackling a particular problem or challenge](#). Understanding the data that exists both within an organisation and beyond can help to build alignment, clarify which data assets should be shared, and help to identify the data stewards, innovators and influencers that can help address a specific social, economic or environmental problem.
- **Improve data discovery.** Cataloging data can help to understand the extent of the data that your project or organisation manages, uses or publishes. Publishing a data inventory under an open licence and using open metadata standards can help others to find, access and use the datasets that your project or organisation may be able to share. A data portal is an example of a public data inventory.
- **Understand a particular ecosystem.** A data inventory can be created as part of a data landscaping² exercise, to engage with a particular ecosystem of actors³ and identify datasets used and shared by stakeholders to

² (2020) Open Data Institute, 'Data Challenge Prizes for Health: a Playbook'. Accessed December 2020.

³ (2019) Open Data Institute, 'Data ecosystems mapping methodology and training tools'. Accessed December 2020.

help solve particular problems. Understanding a data ecosystem can help identify organisations, communities and people that will potentially benefit from available data sources. An inventory can help to identify or prioritise which datasets could be published for others to reuse, and to give feedback on which additional datasets need to be collected or shared to enhance product development.

- **Build trust.** Sustainable access to data supports data ecosystems, giving organisations the confidence to invest in data as a means to develop new products, services and research. A public data inventory can also be used to increase transparency around the data your organisation or project collects, either directly or through third-parties. This can help to build trust around how a project or organisation uses data, and so increase use and investments.
- **Increase collaboration.** In order to address a specific problem, a project might need to collect data from multiple sources across stakeholders on a particular subject. Designing a data catalog with open standards can make the process of collecting, structuring and integrating data from different sources more efficient. A research data repository that contains datasets from multiple stakeholders is an example of this.
- **Improve governance of data assets.** The process of compiling and managing a data inventory can help to take stock of the data that a project or organisation is managing. Creating an inventory is often the first step in improving the governance of data. A catalog can provide a place to record who is accountable for collecting and managing data as an asset as well as identify duplicated data collection processes or datasets.

- **Create a legal record and protect personal data.** An inventory can provide a legal record of the data that an organisation manages. This might be done for compliance reasons when managing personal data, or to maintain a list of third-party datasets your organisation accesses and the licensing and data sharing agreements which govern their use.

Creating and maintaining a data inventory is an important step in treating data as an asset, so that it is accessed, used and shared in ways that will help to maximise value, and meet the needs of a project, organisation, sector or society.

Steps to building a data inventory

A data inventory can fulfil several different functions, such as facilitating data discovery, enabling the access, use and sharing of data, supporting decision making or providing oversight for regulatory compliance. This means that cataloging data can support a variety of user demands that need to be balanced while planning the inventory design.

1) Plan the inventory

Decide on the purpose, scope and granularity of the data inventory.

- **Consider the inventory's purpose(s).** This will help define the type and scope of information you want to collect. For example;
 - If you want to increase data quality and availability, you may need to focus on the problem you are trying to tackle, look at standards or other stakeholder's inventories.
 - If you want to improve data management, you'll need to collect enough technical information about the data you are

collecting to be able to understand how data is stored and formatted.

- If you want to identify business opportunities, you'll need to know about your rights to access and reuse the data, the costs of maintaining the data and its value.
- For compliance reasons you may need to identify which datasets contain sensitive or personal information.
- **Outline your definition of 'data' and decide what to include.** There are many ways of defining 'data' and so it's important to start with a common understanding of the scope of your inventory. When defining the data, you might also identify any personal or commercially sensitive information, or third party data.

For example, do you want to collect information on physical or paper-based assets? Is it for all information assets (for example reports, data visualisations and analyses) or just 'datasets'? Do you want to include data that was produced by third parties but forms part of your workflows? The following dataset definition may be a useful starting point:

“A dataset is a collection of data that relates to a common topic or was curated for a common purpose. A dataset has a consistent standard in terms of its format and structure. A dataset can contain 'raw data', analysed results or derived information⁴.

⁴ Derived information is new information or datasets that are created from existing data

- **Consider the inventory's sustainability and governance.** Sustainable access to data is critical to ensuring that data remains findable, accessible, interoperable and reusable (FAIR) and is safeguarded in the long term. Allowing a wider group of people to use the data over time maximises its utility. To provide sustainable access to data, data inventories need to be kept up to date, and the responsibility for collecting the data and managing the inventory needs to be clear. You can ensure this by identifying the agencies or data stewards responsible as well as those who will be in charge of quality checks and monitoring of the data inventory. Once governance is defined, we recommend adding inventory updates and validation processes into existing workflows, for example, when a project is opened or closed, the project manager or data steward should register new or updated information to the inventory. You should also consider where this inventory will be hosted and how people will access it.
- **Identify and mitigate potential risks.** If the scope of your inventory includes third party datasets, personal or commercially sensitive information, you will need to assess any potential risks of storing, using and sharing this data and consider how to reduce potential harm to individuals or communities.

- **Engage users.** Consider and engage those that will use the inventory for its intended purpose. It is important to ensure the data will suit the needs of your audience, that might differ for each type of data to be managed. Engaging with colleagues, partner organizations, public institutions and external users may help you understand what information is important to them and how they might search for it so that you can structure and populate your inventory in the most useful way. When designing a data inventory you might want to build [personas](#) from your data community or [map your data ecosystem](#) through a workshop to understand stakeholders, their data needs and intentions to use your inventory, and find opportunities to catalog data based on their and your needs.
- **Decide what level of detail you need.** Do you want to record information about each single data asset, or a higher-level view of dataset collections grouped by subject, timeframe or creator? It is better to collect detailed information and then catalog the information for different audiences to increase discoverability and use of data assets.

2) Decide attributes to collect

Having considered your inventory's purpose, decide on a set of attributes that you will use to describe the data assets – this is called the 'metadata'. The attributes you ultimately choose to collect will be influenced by the purpose and intended users of your inventory, as defined during the planning stage.

We recommend the following metadata as a starting point:

Attribute	Description
ID	Unique identifier for the dataset
Title	The name of the data asset
Description	A description of the data asset
Purpose	Why was the data collected or produced?
Data creator	Who created the data?
Data manager/owner	Who manages the data?
Subject/keywords	What subjects/topics does this dataset cover? Adding tags to the dataset will help discovery for users searching for this data. It is recommended to use a controlled vocabulary for this attribute (and others where possible) to improve future search and data-linking potential, for example finding related datasets.
Language	In which language(s) is the data available?
Location	Where is the data located or stored?
Creation date	When was the data created?
Update frequency	How often is the data updated?
Type	What type of data is it? Text, numbers, statistics, images, a database?
Format	What format is the data in? For example, XLS, XSLX, CSV, JPEG, SQL DB, ODS, JSON, GEOJSON
Rights and restrictions	What are the access and usage rights and restrictions? If you are publishing the data, what can users do with the data? Include a link to the relevant licence for use of the data (for example a Creative Commons or bespoke licence).

You can refine the suggestions above to include more attributes based on the purpose of the inventory and user needs. Examples of additional attributes you may wish to include are:

- Spatial distribution of each dataset (i.e. which geographic areas it covers)
- What time period it covers
- How widely it is currently shared and/or with whom
- Whether the dataset contains personal or commercially sensitive information
- Information about the quality of the data
- Additional tags to describe the dataset

We have provided template examples of other attributes in the 'additional resources' section below.

Describing and structuring metadata in a standard way can make it easier for people and organisations to find, access, share and use data. There are a number of standards that can be used to describe metadata that will make data assets discoverable. Where possible we recommend using open standards, such as the [Data Catalog Vocabulary \(DCAT\)](#), and you can find other open standards in [VEST/GODAN'S agrisemantics standards](#) or the [ODI's Open Standards Handbook](#). We have listed other examples of metadata standards for different purposes in the 'additional resources' section below.

When cataloging datasets, we highly recommend publishing the license being used so that the legal terms to access, use and share that data are clear for all users. When deciding how to provide access to data, you might consider the specific context, such as the type of data,

whether it contains any personal or sensitive information, the purpose of sharing and who you intend to share it with.

Data licenses can take different forms. An open data license may be used when publishing open data, while a non-open license, or **data sharing agreement**, allows data to be shared with specific groups under specific conditions. A clear and appropriate licence can help maximise the value of data by supporting the development of new products, services and research. If you decide to publish a dataset under an open license, we recommend the Data Sharing Toolkit guide 'How to choose an open data licence'. You can also use the Module 7 guide to designing data sharing agreements when you have identified which agreement you will use.

3) Populate the inventory

Depending on the volume of data you'll be adding to the inventory, and the availability of existing sources of information, you might need to use some of the following techniques to help collect the data for your inventory:

- **Delegate to existing data stewards**
 - Identify data stewards and ask them to complete the metadata directly for the datasets for which they are accountable and responsible.
 - Build on information collected through other activities, audits, and IT documentation.
- **Conduct interviews with data stewards and product managers**
 - Individual or group interviews can help to understand context, for example, by using coaching techniques to help interviewees identify the data they create and use to inform decision making.

- Collect all detailed information about the data as well as any extra information that might be helpful to users, such as recommendations on how to properly use the data assets, any caveats or limitations on the data and the original purpose for which the data was collected.
- **Conduct user surveys**
 - User surveys can help to understand needs and constraints data stakeholders have when accessing or using data.
 - Electronic or web-based questionnaires help to reach many respondents at once.
 - Include commentary and guidance with questionnaires so information is collected following the same criteria.
 - Be aware that sometimes questionnaires have the disadvantage of achieving low rates of return, so try to be concise when designing questions in order to receive feedback you need.

4) Publishing the inventory so others can find it

Publish your inventory so that it can be found, accessed and used by others. In the appendix to this guide, we provide a template of a data inventory with key information.

Publishing can mean making the inventory available to colleagues within your organisation or working on your project, or making it available online for external users to discover what data you hold.

Publishing an inventory does not imply that all the datasets it contains are necessarily available to others. However, a data inventory provides transparency about the data an organisation or project is collecting and using in its products and services, and to inform decisions.

The choice of format for your inventory is likely to be informed by the amount of metadata you are collecting and the overall purpose of your inventory, for example:

- For smaller projects, the inventory could be stored in a spreadsheet that is periodically updated.
- For larger quantities of information, a database might be more suitable. If the aim of the inventory is to improve discovery and data sharing, you might use the inventory as the basis of an online database, which could allow users to search, browse and link to different data. The [VEST/GODAN agrisemantics registry](#) is an example of what such an online inventory might look like.

Publishing an inventory under an open licence can help users find the data they need.

Data should be [as open as possible](#), while still protecting people and communities from harmful impacts. Some data can be published as open data, for anyone to access, use and share. Other data might only be available via a data sharing agreement or governed by another data access model, such as a [data institution](#).

5) Updating the inventory

Plan updates to your data inventory. A sustainable data inventory contributes to ensuring that data access remains findable, accessible, interoperable, reusable (FAIR) and safeguarded in the long term.

- **Plan updates**

- For anything other than a one-off audit, you will also need to plan how to keep the data inventory up to date. Consider planning these updates when designing the data management plan.
- One way to do this is to ensure that the data inventory is embedded into internal data governance and management processes.
- Continuous engagement with stakeholders is recommended in order to keep the inventory up to date. If you publish datasets in a portal, consider including a mechanism for user feedback.

- **Create automated processes**

- Automating processes to complete or update the metadata in a data inventory is an option when it is likely the metadata will change regularly or the inventory will need to be accessed in the long term.
- Introduce protocols so each time a document, dataset or database is added to a content management system, the user is prompted to complete basic metadata that is automatically added to the inventory.

Further resources

Data inventories and assets

- [What is a dataset?](#) – a blog post reviewing dataset definitions
- [UK Data Service](#) – data inventories for research centres: [Data Inventory Template](#) and [guide to metadata standards](#)
- [Data Asset Framework](#) – a set of methods to identify, locate, describe and assess how organisations are managing research data assets: [Data Asset Framework Methodology](#)
- [Global Open Data for Agriculture and Nutrition \(GODAN\)](#) – the [Agriculture Open Up Guide](#) developed by GODAN and Open Data Charter is a data inventory designed to identify government’s datasets needed to enhance agricultural production.
- [GovEx Labs](#) – a [data inventory guide](#) with open resources on how to catalog data

Standards for metadata


- [DCAT](#) – the [Data Catalog Vocabulary](#) is an open vocabulary designed to facilitate interoperability between data catalogs published on the Web. This [blog post](#) can help you get started.
- [ODI open standards handbook](#)- a [guidebook](#) to support organisations create, develop and adopt open standards for data
- [Dublin Core](#) – a set of open source vocabulary terms used to describe digital resources: [DCMI Metadata Terms](#)

- **ISO 19115** – this international standard defines the schema required for describing geographic information and services by means of metadata. It provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services. This is a non-open standard.
- Digital Curation Center – a list of **extensions and tools** related to ISO 19115 extensions
- UK’s Geospatial Commission – **a blog with resources** to share geospatial data
- GODAN – a set of **resources of existing vocabularies** for the exchange of data in agriculture
- **UK GEMINI** – UK GEMINI (GEO-spatial Metadata INteroperability iNitiative) is a specification for a set of metadata elements for describing geospatial data resources.

Appendix: Data inventory template

The scope of the data inventory and the metadata attributes of the data assets to be collected will depend on the project’s purpose, the type of data being collected and the user’s needs. As such, in order to give an example of an inventory template, we have taken a particular use case following the ‘Understanding personas in agricultural data’ guide in Module 3.

Alan has received funding to develop an online platform to improve agriculture decision making. He needs to combine data from multiple sources, types and sectors, so a data inventory is helpful to collect secondary research data, geospatial data from multiple sources and additional data assets generated by the project.



Alan
Lead Organisation

"I want to streamline how my partners access, share and publish data"


Alan's organisation has received funding to develop an online platform with the aim of improving agriculture decision making for researchers, policymakers and innovators. He needs to cooperate with various stakeholders involved in collecting and processing data to make it accessible via the platform.

Goals

- Create a Soil Intelligence Service (SIS) allowing stakeholders in multiple jurisdictions to access raw data, research and software code
- Combine soil data with crop, agronomy, and geospatial data using machine learning models to make it more useful
- Develop a sustainable business model based on a mixture of open (free) and subscription-based products/ services
- Create measurable impact which he can document for his funders

Pain points:

- Difficult to discover where all the data needed for the platform is held, and who has rights to the data
- Harmonising data produced using different formats, metadata, and process flows
- Negotiating access to data with different stakeholders, with different sensitivities
- Government data stewards unwilling to share data outside of their jurisdiction
- Tracking downstream use of data assets to report impact/return on investment



[Explore Alan's role in the ecosystem](#)

Key exchanges he is involved with

- Publish openly licensed datasets and supporting code;
- Publish openly licensed software;
- Procurement of data collection services;
- Recipient of funding

Needs

- Access to high quality, standardised data from government, research institutions, and industry eg crop and agronomy
- Shared data management strategy outlining approach to data governance, access, sharing, publication, rights to use, responsibilities, and sustainability
- Buy-in from government data stewards
- Metrics to monitor progress and measure impact over time from data use
- Subcontract soil data collection services from research institutions, to collect and validate data in standardised way

Data collected

- Secondary research data (e.g soil grids)
- Lab data (e.g spectral data)
- Geospatial data (e.g remote sensing covariates)
- Software code

Key learning resources

- Data management plan checklist
- Data business models
- Open standards for data toolkit
- Data inventory guide
- Data sharing agreement guide
- Data ecosystem mapping tool
- VEST Agrisemantics map of standards
- Procurement guide

[Tools for a lead organisation](#)

1. Plan the inventory

- **Purpose:** Alan's main goal is to create a Soil Intelligence Service and increase data availability by combining soil data with crop, agronomy and geospatial data.
- **Data definition:** Alan plans to collect secondary research data, lab data, geospatial data and software code.
- **Sustainability and governance:** Alan should identify data stewards who will be in charge of managing and providing sustainable access to the data collected by the project.
- **Assess risk:** In order to understand the risks related to the data this project might collect, Alan will refer to the Data Sharing Toolkit guide 'Sharing agricultural data: managing risk to minimize harm'.
- **User engagement:** Alan also assess the data ecosystem, and look at the standards used to catalog this type of data. By looking at GODAN's agrisemantics standard, he identifies the [International Consortium for Agricultural Systems Applications \(ICASA\)](#) as helpful in standardizing the data specifically collected for crop growth modeling. Another important source for the project is ISRIC's SoilGrid data portal and [soil data standardization manual](#).
- **Level of detail:** Given that some agriculture data is being collected by other organizations, Alan's project is planning to collect data assets from these sources and host additional datasets that might be created by the project. The following fields to support data cataloging are based on [GODAN's Agriculture open up guide](#):

Field	Data Asset	Description	Location	License
Land Use	Land use dataset	Datasets describing cultivated areas around the world	http://www.earthstat.org/	CC-BY 4.0 License
Soil	Soil maps Soil classification dataset	Soil data describing characteristics and classes provided by ISRIC's world soil information FAO's harmonized soil dataset	https://soilgrids.org/ http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/	CC-BY 4.0 License
Access to water	Water sources map and location datasets	EU's open data portal datasets related to water sources to support agricultural projects	https://data.europa.eu/euodp/en/data/dataset/DAT-163-en	CC-BY 4.0 License

2. Decide on dataset’s metadata

When collecting and generating data assets, Alan will need to describe data using metadata attributes.

For the purpose of the project, he subcontracted a university to collect data related to soil quality in X region in order to support the decision making of small farmers and SMEs. He needs to decide which metadata he will be using to describe this data:

Metadata	Definition	Unit of measure	Example
dataset_ID	Unique identifier	number	
dataset_title	Name of the data asset	text	Soil quality X region
dataset_description	Description of the data asset	text	Map providing soil quality in X region
dataset_purpose	Why was the data collected or produced?	text	Provide data to farmers and SMEs on soil quality in the X region
dataset_creator	Who created the data?	text	Name of University
dataset_manager	Who manages the data?	text	Alan
dataset_keywords	What subjects/ topics does this dataset cover?	text	soil, soil management, land, environment
dataset_language	In which language(s) is the data available?	text	EN
dataset_location	Where is the data located or stored?	text	URL
dataset_creation	When was the data created?	date (yyyy-mm-dd)	2019-11-30
dataset_freq	How often is the data updated?	text	monthly

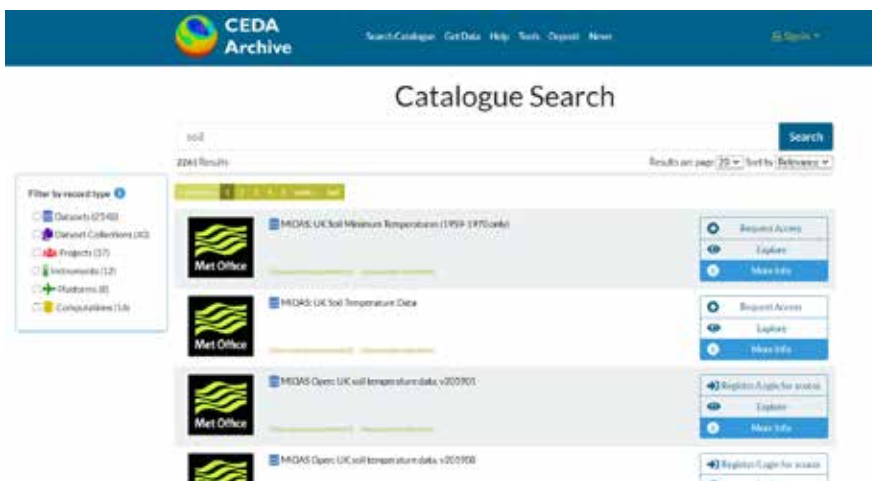
Metadata	Definition	Unit of measure	Example
dataset_type	What type of data is it?	text	images
dataset_format	What format is the data in?	text	GEOJSON
dataset_license	What are the access and usage rights and restrictions?	text	Link to the license
dataset_spatial	What region/s or countries does the dataset cover?	text	X region
dataset_time_period	What time period is covered?	date (yyyy-mm-dd/yyyy-mm-dd))	2019-11-30/ 2020-11-30
dataset_privacy	Does the dataset contain any personal or commercially sensitive information?	value 0:the data does not contain any personal or commercial information 1: the data contains commercially sensitive information 2: the data contains personal information	0
dataset_quality	What is the quality of the dataset? Is it updated?	value 0: the data quality is poor and not updated 1: the data quality and update is adequate 2: the data is good and updated on a regular basis	2

3. Publish the inventory

Given that the project’s purpose is to increase access to data for decision making, Alan considers publishing this inventory online, in a data portal or directly on the project’s website. Some examples on how to publish the results of cataloging data online include:

- Designing catalog search in the website, so that users can explore the different data assets from multiple data sources, request access if needed and have the dataset’s description (ex. CEDA’s Archive catalog search)
- Publishing your data inventory as a table or spreadsheet with a set of defined tags/fields based on the specific fields that your data assets cover (ex. Philadelphia’s data catalog)
- Publishing your inventory in a data portal, containing the associated metadata and download details (ex. USDA’s agricultural data catalog)

Here are some visual examples of published inventories:



Source: [CEDA Archive catalogue on soil data](#)

Data Sharing Toolkit



ACKNOWLEDGEMENTS

This document was authored by the Open Data Institute and CABI as part of a Bill & Melinda Gates Foundation funded investment.

The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation or CABI.

datasharingtoolkit.org

DOI: [10.21955/gatesopenres.1116749.1](https://doi.org/10.21955/gatesopenres.1116749.1)

This is an updated version of a document previously published on Gates Open Research.

[10.21955/gatesopenres.1114885.1](https://doi.org/10.21955/gatesopenres.1114885.1)

cabi.org | theodi.org | gatesfoundation.org

 **CABI** Data Sharing Toolkit



BILL & MELINDA
GATES *foundation*



Except where otherwise noted, content on this site is licensed under a Creative Commons Attribution 4.0 International license.