Data Sharing Toolkit

Module 7

Developing a data management plan

Guide

Developing a data management plan

Contents

Purpose of this guide

This guide is intended to support grantees to facilitate making data as findable, accessible, interoperable and reusable (FAIR) as possible, by creating a management plan for using and sharing datasets within investments funded by the Bill & Melinda Gates Foundation.

It provides a checklist and guidance to create a data management plan, guidelines on how to manage data as an asset, and Resources for data management and planning.

The guidance is intended to support the following outcomes:

- Data is made FAIR[1] – Findable, Accessible, Interoperable and Reusable – so that value may be extracted from it by or in collaboration with others.
- Data is made as open as possible, for anyone to access, use and share.
- Any risks of personal data are assessed and harmful impacts are minimised.
- Sustainable access to data enabled and ensured.
- Data is stored securely, safeguarded and not lost or corrupted.

However, there is no one-size-fits-all approach to managing data, and the nature of the data, the specifics of the project and its intended use will need to be taken into account when considering the appropriate course of action and the importance of each section of the checklist.

[1] Wilkinson, M. D. et al. (2016), *The FAIR Guiding Principles for scientific data management and stewardship*. Scientific Data 3:160018 doi: 10.1038/sdata.2016.18

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:

- When creating a proposal to inform the project design, budget, and estimate the effort and cost of managing data
- At the beginning of an investment to consider and document the types of data a project will need to access, use and share
- When establishing the project itself to ensure that all relevant aspects of data management have been considered and that people, processes and resources are in place
- During project implementation as a tool to help locate, manage, use and share data within investments
- When completing a project and making final decisions on archiving or publishing data

This guide is not legal advice. If you are uncertain, seek guidance from a legal professional.

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.

A data management plan supports funded programs in documenting the landscape, collection, storage, access and sharing of data within a project and ensure that this data is 'as FAIR and open as possible'[2] and managed in a way that minimises harmful impacts.

[2] Open Data Institute 2020, Creating FAIR and open data ecosystems for agricultural programmes, https://gatesopenresearch.org/documents/2-42 Accessed November 2020

A data management plan documents the technical processes needed to manage data, and the resources required to support them. Helping to assess and mitigate risks, ensure sustainable access through planning and resourcing, and understand the potential impacts (positive and negative) of data sharing, a management plan supports the effective delivery of a project and also increases trust among delivery partners.

The guide presents a list of key considerations when putting together a data management plan and guidelines on how to manage data as an asset. The project, the nature of the data to be managed, and the landscape in which it will be collected, shared and used will determine the level of importance given to each of the items included here, and the course of action will vary accordingly.

A data management plan should consider How the data will be stored and accessed

- How the data will be described, including naming conventions and metadata
- How the integrity of the data will be maintained over time
- How the security of the data will be maintained over time
- How the data will be shared at the end of the project
- What resources are required.

The checklist

Actions	Activities	Is it considered in your data management plan?	
		Yes	No
Putting together a data management plan	Decide on the data governance model		
	Assign roles and responsibilities		
	Ensure sustainable access to data		
	Consider resources		
	Plan for storage and security of data		
	Identify training needs		
Managing data as an asset	Define the data		
	Identify users of the data		
	Choose an appropriate format		
	Consider personal data		
	Consider data ethics		
	Create a data inventory		
	Plan for data processing		

This checklist should be used as guidance only when creating a management plan and managing data as an asset and should not be thought to represent any order of importance, nor suggest a linear workflow to be followed when creating a management plan.

Creating a data management plan

This section includes a list of key considerations when putting together a data management plan and guidelines on how to manage data as an asset.

Designing a data management plan includes outlining the technical processes, roles and governance to effectively manage data, and the resources required to support the collection, access and use of data.

Creating a data management plan should, ideally, be a collaborative, iterative, process involving all stakeholders in a project, including funders. This guide can be used alongside specific instructions on data management provided by funders and with their templates for data management plans, where they exist.

Putting together a data management plan

Decide on the data governance model

Data governance, or stewardship, involves collecting, maintaining and sharing data. In particular, it involves determining who has access to data, for what purpose and to whose benefit.

How data is stewarded affects who can access it, what it can be used for and how it may bring benefit or cause harm. Data can be stewarded by the organization that collected it or by a third party.

Increasingly third party organizations are stewarding data on behalf of others. These types of organization come in different shapes and sizes and are referred to by different names such as data institutions[3] and data collaboratives.[4]

[3] Open Data Institute (2020) "What do we mean by data institutions?" https://theodi.org/article/what-do-we-mean-by-data-institutions/. Accessed 13 Jan. 2021.

[4] TheGovLab, 'Data Collaboratives', Accessed September 2020, https://datacollaboratives.org/

These parties play a number of roles, including providing access to data on behalf of others, combining or linking data from different sources, and developing and maintaining common data infrastructure. Using a third party can help foster trust with other actors in the data ecosystem.[5]

The method by which you choose to steward data will be informed by 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.

Assign roles and responsibilities

Within an organisation or project, data management responsibilities should be assigned to data stewards, named individuals acting in a particular role or particular departments within an organisation. The data steward is likely to be the business or policy lead that the data relates to. They are ultimately accountable for responsible data management, although they might delegate some tasks to the IT specialists within the company for storage, backup and security, for example.

For grant-funded projects, it is particularly important to make clear at the outset who will be accountable for managing the project's data after the funded work has been completed. This might involve a transfer of responsibility to another organisation or to the appropriate government body, who will need to be involved in any planning as early as possible.

Be clear on how these responsibilities are covered within your organization's resources and which are funded exclusively by the grant, especially in the case where the need for the role or activity extends beyond the lifetime of the investment.

[5] (2020), Open Data Institute, 'Designing sustainable data institutions'. Accessed June 2020. https://theodi.org/article/designing-sustainable-data-institutions-paper/

Roles in a data project may include

- **Data steward**: project or departmental lead that the data relates to and ultimately accountable for responsible data management. This person should be in a senior role.
- **Data protection officer/legal advisor:** responsible for giving legal advice on data protection compliance matters, legal guidance on how to ensure digital rights, and advice on data access agreements and legal terms for data use.
- **Data scientist:** data expert in charge of processing, exploitation and analysis of data.
- **Data architect:** responsible for designing and implementing technical specifications to collect, share and use data**.**
- **IT specialist/cybersecurity:** in charge of securely storing data and ensuring its integrity.

Identify who is responsible for

- Short-term management of the data, including protocols for data capture, within the duration of the funded project
- Long-term management of the data after the project is finished
- Creating a data inventory, cataloging data and ensuring data quality
- Monitoring ethical use of data
- Monitoring the security and integrity of the data
- Data protection and compliance where personal data is to be collected or processed, ensuring that the necessary consents have been obtained and recorded, and that digital rights are respected
- Other legal responsibilities including the management of sharing agreements and use of third party data
- Analysis and exploitation of the data, within the project – including collaboration with data users and other data providers
- Publication and dissemination of data collected if the data is open or shared
- Provision of storage and backup facilities, during and after the project

Sustainable access to data

Ensuring data remains accessible over time maximises the utility of the data by making it possible for the greatest number of people to benefit from it by making better decisions and reducing risk.

The individual with responsibility for long-term management of the data after the project is finished will need to plan and resource the collection, access and use of data over time.

Consider the resources required to manage the data

The data steward will need to plan how to source budget and resources necessary, for:

- **Storage –** Short-term storing and processing data during the project and longer-term, for storing, publishing, ensuring access is maintained and retaining rights to it after the project has ended
- **Collection and use**
  - Specialist equipment or applications, where necessary, to collect data, eg electronic tablets with data collection apps for field work
  - Cleaning and formatting data
  - Data quality control
  - Equipment or computer facilities for data processing
  - Software, and any associated licences (or support contracts for open source software) to use data

More guidance can be found in the Data Sharing Toolkit guide ‘Ensuring sustainable access to data’.

Describe how data will be collected and transferred to secure storage

As part of ensuring sustainable access to data, it is important it can be shared and accessed by other organizations where appropriate. The following considerations are useful for securely storing, using and sharing data.

- **How might calibration or other contextual data be collected? How will this data be associated with the raw data to which it pertains?**
  - If data collection is to be subcontracted, ensure that licensing, intellectual property or personal data handling requirements are addressed in the agreement made with the contractor.
  - In the case of manual recording (eg in a lab book), describe how the data will be transferred to an electronic format and the quality controls that will be put in place to guard against transcription errors. Data should be transferred to a secure, backed-up, electronic format as soon as possible after it has been collected.
- **If data is to be collected using an app running on a tablet or mobile device, how will data be transferred to central storage?**
  - This might be as a ‘drip feed’, transferring data as it is collected by means of a wifi or mobile phone network, or in bulk on return to a lab or office, using wifi or a cable connection.
  - If a wifi or mobile network connection is used, the collection device should be able to store data locally, for download later, in the event the network is not available.

If the data collection app is bespoke to the project, how will it be built, tested and maintained?

- **If the data contains personal or sensitive information, how will this be kept secure during collection, transport and storage. How will access to it be managed, and protected from unauthorised persons?**
  - If personal or sensitive data is held or processed, it is essential to ensure that data cannot be seen or taken by unauthorised persons. Good IT security practice for the protection of data and systems should be followed.
  - The Data Sharing Toolkit guide 'Managing risk when handling personal data' provides guidelines on how to manage personal data and assess risks.

Decide how to keep data secure and guard against loss or corruption

Data must be secured against loss, usually by means of a backup mechanism. Data should be copied to another store, preferably at a separate site to ensure preservation in the case of equipment failure, fire or flood at the primary site. This might be a service provided by an organization's IT department or by a data repository, if one is used.

The frequency that backups are made should depend on the data's rate of collection. For example, where new data is created once a week, back ups should be taken weekly rather than daily, which would be too often, or monthly, which would not be enough.

Ensure that there is a plan to test the retrieval of data from the backups. The first time to test this is not when you need to do so for real.

For large datasets in particular, consider using a hashing or checksum mechanism to detect corruption of the data after it has been retrieved from storage or transmission.[6]

Decide how to store the data during the lifetime of the project

Data might be collected onto paper, a tablet, a laptop or a desktop computer, but it should be transferred to another medium that poses a lower risk of loss or corruption as soon as possible. This would ideally be a central data store, such as a network drive supplied by an institution's IT department , cloud storage, such as those provided by Amazon, Google, Apple or Microsoft, or a repository such as the Open Science Framework (OSF)[7] or similar site appropriate to the subject domain.

All of these services will include backup and disaster recovery services, and will deal with data security to a greater or lesser extent, but check that what they provide meets your particular needs. Consider:

- **What volume of storage is required?**
  - Make an estimate of the rate at which the collected volume of data will grow, and ensure that there is provision, both in terms of funds and available resources, to accommodate this storage growth. Take into account raw data and any derivatives from it, such as the outputs from analysis or modelling activities.
- **Are particular storage facilities needed to suit the format of the data?**
  - If the data is in a specialist format – GIS layers, for example – specialist data stores might be needed in order to make the data accessible and usable to those that need it.

[6] A *hash function* or *checksum*, such as MD5, is used to create a short numerical sequence (a *hash code*) that is unique to the data concerned. If, after retrieval or transmission, the same hash function is applied to the data and produces a different hash code, then the data has changed in some way. A single character change in even a very large dataset will give rise to a change in the hash code.

[7] https://osf.io/

Design a data retention schedule

Data retention refers to how long a dataset is, or needs to be, accessed, used and shared. You need to consider retention periods for each type of data created by your project. This includes deciding what data should not be accessible when the project ends. For instance, best practice data protection legislation states that personal data should not be stored 'longer than is necessary for the purposes for which the personal data are processed'.[8]

- **How will the data be stored and maintained in the long term, after the funded project has been completed?**
  - If you have used a repository to store the working data, like the OSF mentioned above, you might keep data there in the long term. If you have used another mechanism, such as institutional storage, you will almost certainly need to make provision for storage of, and access to, the data after the project is complete.
  - The Gates Open Research platform provides guidance on preparing data for publication and a non-exhaustive list of repositories approved by the foundation.[9] It also describes the desirable properties of a repository, which should enable access to the dataset, ensure it's persistence and stability, and enable searching and retrieval.
  - The journal PLOS ONE recommends[10] repositories suitable for data associated with particular fields of study, and the journal Scientific Data makes similar recommendations.[11]

[8] (2020) Information Commissioner's Office UK, 'Guide to the General Data Protection Regulations, Principle (e) – Storage Limitation'. Accessed December 2020. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/storage-limitation/

[9] https://gatesopenresearch.org/for-authors/data-guidelines

[10] http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories

[11] https://www.nature.com/sdata/policies/repositories

- **How will data be selected for long-term storage and for wider dissemination?**
  - It might not be appropriate to maintain all data produced by a project. Selection criteria for long-term storage, and for dissemination beyond the project itself should be established at the outset, along with clear retention periods for each category of data generated by the project.

Identify training needs

When planning on how to manage your data project, make sure that your organization has the skills to effectively access, share and use data and identify any additional training needed to implement your plan. One way to identify training needs to effectively manage your data project is to implement a users survey.

We recommend looking at the eLearning modules that the Open Data Institute and CABI have produced to support FAIR and safeguarded data in investments made by the Bill & Melinda Gates Foundation to help grantees and grant officers access, use and share data in grants, while minimising harmful impacts.

Managing data as an asset

Creating value from data is only sustainable when it is governed, resourced and stored properly, in a way that safeguards against harm.

This section provides guidelines that help to plan how data will be collected, used and shared, and which technical specifications need to be in place to manage data as an asset.

Define the data

Be clear on the specifics of the data you will manage and its intended use both within the project and by other stakeholders.

- **Outline your definition of 'data' and decide what to include**
  - There are many ways of defining 'data' and so it's important to establish a common understanding of the scope of your inventory.
  - Identify whether you are collecting personal or third-party data and what intellectual property rights need to be considered. The guide 'managing risk to minimise harmful impacts' in the Data Sharing Toolkit can help you with this assessment.
- **Catalogue your data in an inventory**
  - We recommend cataloging data to define what data is being collected and shared, and to describe the data assets (the metadata) of the project to ensure discoverability and use.
- **Does "data" include the outputs from analysis or modelling work? If so, what is the nature of this data?**
  - We recommend including metadata to highlight the source of these outputs, what they include and any limitations with them as for other types of data.

The Data Sharing Toolkit guide 'how to create a data inventory' explains more about defining and cataloging.

Identify the users of the data

The data's key stakeholders will influence a lot of decisions around how the data is defined, managed and made available to others.

It is important to ensure from the outset that the data will suit the needs of users throughout the data ecosystem. Visually mapping out the actors that will access, use and share the data, as described in the ODI's guide to mapping data ecosystems can help to identify users and potential users.

The users will almost certainly be different for each type of data to be managed. For example, the audience for raw data will be different from that for data or insights that are generated through any analysis based on that raw data.

Stakeholders in your data ecosystem may include:

- **Data stewards:** Responsible for collecting, managing and ensuring access to a dataset; may include provision of infrastructure, data governance, etc
- **People or organisations**: People or organisations the data is about or who are impacted by its use.
- **Contributors:** People or organisations who contribute to or help curate a dataset; they may do so knowingly, using tools and frameworks provided by a data steward, or unknowingly through their use of a service.
- **Regulators:** Create the policies and legislative frameworks within which others operate.
- **Intermediaries:** Provide value-added services that wrap, host or enrich a dataset.
- **Aggregators:** Package together datasets from many sources and are a type of intermediary.
- **Creators (or reusers)**: Use data to create information, in the form of products and services, analyses and insights, or stories and visualisations.

- **Beneficiaries:** People or organisations that benefit from the data ecosystem by making better informed decisions through the use of products and services, along with their own experience and understanding.
- **Researchers:** A type of creator who uses data for research purposes.
- **Policymakers:** Create principles and measures to generate outcomes.

It helps to think broadly, identify whether likely users include individuals, specific groups or the wider public. Consider whether your users span across the public, private or third sectors. Engaging the user community and developing personas to represent real people in your data ecosystem can help to identify audience needs, pain points, motivations and goals.[12]

Choose appropriate formats

- **In what format will the data be collected?**
  - Data might be collected by hand-writing into a lab book, by means of an application running on an electronic tablet or by capturing the output from a machine onto a computer.
  - For electronic capture, consider in what format will the data be represented (for example, MS Excel, csv, tab-separated text file, json, geojson).
- **In what format will the data be stored and made available to others?**
  - Using common formats for publishing data can facilitate data discovery and exchange. The ODI provides guidance on choosing open standards.[13] You can also consider UK Data Services recommendations on appropriate data formats[14] or this catalogue from FAIRsharing.org.[15]

[12] (2018), Fiona Smith, Leigh Dodds, Pauline L'Henaff, Charlotte Day, Ruthie Musker, Martin Parr, 'Understanding personas in agricultural ecosystems'. Accessed June 2020. https://gatesopenresearch.org/documents/2-43

[13] http://standards.theodi.org/find-existing-standards/how-to-choose-an-open-standard/

[14] https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

[15] https://fairsharing.org/standards/

- Many fields of study have already designed an open standard for collecting sectoral data (see for example GODAN's agri-food standards), and the use of these formats might help other stakeholders to process and use this data.
- When your project collects data that can be shared with others, avoid proprietary formats associated with particular equipment or software packages. In general, try to use the simplest, open format that will represent the data so your stakeholders can access and use this data.
- If data must be collected in a proprietary format, convert it to an open format as soon as possible, while retaining and storing the original, of course.
- It is wise to avoid, where possible, compression or encryption. These processes might only be reversible by those with access to specific software, which might no longer be available. If data must be compressed (to manage storage space) or encrypted (for privacy issues), then avoid proprietary formats and favour open or well-known formats such as GZIP[16] for compression and Open PGP[17] for encryption in transit.

[16] https://www.gnu.org/software/gzip/

[17] https://www.openpgp.org/

Consider personal data

When defining what data your project will be managing, it is important to think about data about people, and take into consideration both the opportunities and risks when collecting, using and sharing this type of data. Data protection regulations across the world are designed to enable personal data to be used while minimising the risk of harmful impacts. These regulations define the lawful basis for collecting and using this data, the rights of the data subject, and the liabilities and penalties that your organization

needs to consider when defining the data management plan. For more information on data protection considerations, refer to the Data Sharing Toolkit guide ‘minimising risk when handling personal data’.

- **Does any of the data to be collected or processed contain personal information?**
  - When collecting or using personal data, you might consider assessing risks to minimise harmful impacts to people, communities or other organizations.
  - If personal data needs to be managed, there will be a need to comply with local legislation on the storage and processing of such data.
- **When collecting personal data, how will you minimise harm to individuals?**
  - Consider data minimisation by design. If you don’t need personal details from individuals or commercial information, don’t collect them.
  - Anonymisation techniques can be applied to reduce the risks of re-identification and possible harm resulting from sharing data about people.
  - Guidance on techniques and ways of reducing risk can be found in the Data Sharing Toolkit guide ‘anonymising data in agriculture’.

Check the Data Sharing Toolkit for further recommendations on protecting personal data and minimising risk when handling personal data.

Identify third-party data

Ensure that the rights to access, use and share third-party data are clear and established in explicit licenses and agreements and that there is a full understanding of relevant legal restrictions or permissions.

- **Is any of the data to be managed within the project from a third party? If so, are there restrictions on its use or retention?**
  - When managing data from other sources, you need to understand the legal terms on how data from third-party sources can be accessed, used and shared.
  - Restrictions on the use of third-party data and/or derivatives of it should be considered.
  - Some guidance on how to access, use and share data from third-party sources can be found in the Data Sharing Toolkit eLearning module ‘reusing data from third party sources’. You can also refer to the guide to data sharing agreements and the ‘considering data rights and permissions in investments’ guide.

Consider intellectual property

By default, the data creator holds exclusive rights to use the data, so others must seek or be given the rights to use the data themselves. In these instances, the rights to access, use and share the data must be considered in order to ensure the permissions acquired from the data creator will facilitate its onward licensing and use. This is important if your project is creating new data.

- **Which organisation is the steward for the data?**
  - Understand which organisation or individual has control over the governance, sharing and publication of data. Note that how this

responsibility extends or changes beyond the life of the project should be agreed and made clear before the data is collected. This should be considered alongside considerations about governance from the creating a data management plan section of this guide.

- **Describe any restrictions for data use or sharing**
  - For example, the Gates Foundation Information Sharing Approach[18] makes a presumption that data will be fully disclosed unless there is a legitimate reason not to, a list of which reasons are made available, including for safety and security, personal privacy or confidentiality and where legal agreements may be breached).
  - You can also find guidance in the Data Sharing Toolkit eLearning module 'Minimising harmful impacts from data sharing'.
- **How will you license the data?**
  - If you intend for the data to be entirely open, a licence will help make this clear for potential users. It will also describe any requirements, such as attribution or restrictions on commercial use Where possible, use a standard, well-understood licence, such as one of the Creative Commons licences.[19]
  - Ensure that the licence you apply to the data complies with the criteria under which funding was provided.
  - For more on choosing an appropriate licence for open content, see the Data Sharing Toolkit guide 'how to choose an open data licence' and the eLearning module on 'Sharing data through data licensing'.

[18] https://www.gatesfoundation.org/How-We-Work/General-Information/Information-Sharing-Approach

[19] https://creativecommons.org/share-your-work/

Consider ethical use of data

When managing data to be collected, used and shared in your project, you might reflect on the purpose of using this data and assess any impactful harms or consequences to people, communities or other organizations. Consider using the ODI's Data Ethics Canvas[20] to guide your conversations and actions around data ethics and the responsible use of data.

If the work that generated the data was the subject of an ethical evaluation, for example an assessment by an ethics committee, or was collected according to a published code of conduct, consider also publishing or providing a link to the terms under which the data was collected. This will make the data more reusable by others.

Plan for data sharing, access and reuse

Whether choosing an open data licence or setting a data sharing agreement, grantees will need to plan on how to provide access to data collected by the project, both in the short and long term.

- **How will the data be shared with others?**
  - Data licenses can take different forms, for example, an open data license may be used when publishing data that anyone can access, use or share, and a non open license, or data sharing agreement, for sharing data with specific groups.
  - When choosing or creating a data license it's important to carry out the appropriate due diligence and data privacy assessments to ensure that you, as the publisher, have the rights to share data in the first place.

[20] https://theodi.org/article/data-ethics-canvas/

- **How will others become aware of, find and access the data?**
  - These are functions of a data catalogue or inventory. If the data is to be deposited in a recognised repository such as a data portal, this repository will probably provide a catalogue facility.
  - Ensure that the metadata you use to describe the data is sufficient to allow others to explore and integrate your data with other sources, where appropriate.
  - Ensure that the dataset has an identifier assigned to it. This will make it easy for others to share its location, and to cite the data in their own work.
- **How will you provide access to data?**
  - There are many ways in which your organization can provide access to data.
  - Sharing data with others will be made easier by designing a data inventory and using an open standard for data.

For guidance on how to share data you can also check the Data Sharing Toolkit guide ‘deciding how to provide access to data’ See the guide to data inventories for more on inventory and cataloguing, and the guide to data sharing agreements when planning and designing an agreement.

Approaches to providing access to data include

- **Publishing online.** Try to select a licence as open as possible, while protecting people from harm, so that anyone can access, use and share the data in future.
- **Delegating data stewardship to a third party.** These types of organization come in different shapes and sizes and may be referred to by different names including data institutions and data collaboratives.[21]
- **Sharing the data under contract.** A contract, such as a data sharing agreement, with detailed, binding rules helps everyone be clear on their obligations.
- **Pooling data using a platform.** Collection of data together in one place, such as in data portals, data warehouses and data spaces. Access can be managed via user authentication if the data contains sensitive information.
- **Using technology to support access and protect data.** Examples include privacy enhancing technologies, application programming interfaces (APIs) and the generation of synthetic data.
- **Data visualisation.** Analysing data and presenting it in a visual form to aid understanding or to convey a message, such as using infographics and dashboards.

[21] (2020), Open Data Institute, 'What do we mean by data institutions', Accessed June 2020. https://theodi.org/article/what-do-we-mean-by-data-institutions/, TheGovLab, 'Data Collaboratives', Accessed September 2020, https://datacollaboratives.org/

Create a data inventory

A data inventory can be a useful tool for managing your project's data and it can also 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. We highly recommend creating and publishing a data inventory to increase access to data and use.

For guidance on how to catalogue your data, use the FAIR data toolbox 'how to create a data inventory' guide.

Decide on a naming convention to organise data assets

- **How will you label and organise electronic files?**
  - It is essential to establish a naming convention for stored datasets at the outset and, if appropriate, a naming hierarchy. Without this, managing anything but a few records will become difficult and unreliable. A naming convention – along with appropriate documentation – will help others find data they are looking for.
  - Make file or dataset names predictable. For example it can help to use a consistent stem followed by a date, always in the same format. This allows users to create scripts to collect and update datasets.
  - If you are likely to hold more than one version of a dataset – if it will be corrected or adjusted in some way after collection – consider how best to implement version control and identification.

Decide how to describe the data with metadata and documentation

When collecting and sharing data you might decide on a set of attributes that you want to describe your datasets – this is called the 'metadata'.

The Data Sharing Toolkit guide on data inventories provides guidance on how to describe and catalogue data within the investment.

- **Decide how to name fields within the data, and provide explanatory documentation**
  - Labels in your data should, if possible, be self-explanatory. If it is not possible to make them so, then documentation linked to the dataset should explain each parameter.
- **Describe the data with metadata.**
  - Metadata describes other data. It allows others to understand the content and structure of a dataset, making it easier to find and use.
  - Consider the use of a restricted or controlled vocabulary for some fields – for example their theme, keywords or tags – to ensure that related datasets can be linked and accessed.
- **Describe the metadata scheme to be used.**
  - Many subject areas have existing metadata schema or standards, and these should be used where applicable. The Global Open Data for Agriculture and Nutrition (GODAN) has produced a map of open standards[22], and the Research Data Alliance (RDA) provides guidance on existing metadata standards.[23]
  - If no standard exists, be explicit about exactly what metadata must be attached to each dataset.

[22] http://www.godan.info/documents/map-agri-food-data-standards

[23] http://rd-alliance.github.io/metadata-directory/

- **How will you make sure this data inventory is up to date?**
  - You might need to plan updates to your data inventory and assign this responsibility to the data steward.

Plan for data processing

While it is likely that value will be obtained from data only by processing it in some way (even if for quality control alone) it is also essential that the original raw data is preserved. This allows the raw data to be re-used if necessary.

- **Describe how the raw data will be processed**
  - What quality controls will be applied, for example to detect erroneous values or to mark, or substitute for, missing data?
  - Where appropriate, how will calibration data be applied to the raw data?
  - Will the data be combined with other datasets to provide context?
  - What analytical or statistical methods will be used to extract meaning from the data?
- **Identify specialist facilities needed to process the data.**
  - If the dataset is large or the processing of the data is complex, it might be necessary to use specialist facilities for data processing and analysis.

Criteria to consider when choosing such facilities include

- **The cost of computation and the units used as a basis for costing.** The cost of an on-premises solution would include buying and running hardware, while the cost of a cloud-based solution would be based on computational time used, or the lifetime of a virtual machine.
- **The costs of data storage associated with computation.** This might be temporary storage, for the duration of the computation.
- **The cost of data transfers on and off the computation platform.** For large datasets, these transfer charges can be considerable when compared with the cost of storage and/or computation.

Consider managing analysis software or scripts alongside the data

- **Is management or analysis software to be stored or deposited alongside the data?**
- It might be appropriate to manage bespoke software (for example Python or R scripts) that has been created to manage or analyse data, alongside the data itself.
- Consider storing scripts for data quality checks.
- Manage version control and documentation and use a separate repository, designed specifically for code, such as GitHub.[24]

[24] https://github.com/

Additional resources

FAIR and open data for agriculture tool audit

The Gates Open Research Data Guidelines offer advice on formatting data for publication and on selecting a repository via which to share data: https://gatesopenresearch.org/for-authors/data-guidelines

The Digital Curation Centre provides a range of how-to guides on various aspects of data management: http://www.dcc.ac.uk/resources/how-guides

FAIRsharing.org provide a catalogue of data standards and policies, including those of major funders: https://fairsharing.org/standards/

The Research Data Alliance (RDA) provides guidance on metadata standards and tools, and some illustrative use cases for metadata: http://rd-alliance.github.io/metadata-directory/

The UK Data Service provides detailed guidelines on data storage, backup and security: https://www.ukdataservice.ac.uk/manage-data/store

UK's Information Commissioner's Office guide to plan a data protection impact assessment: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

The International Center for Tropical Agriculture (CIAT) created a data management support pack to help researchers produce high quality data for improving the quality of people's lives. https://ciat.cgiar.org/data-management-support-pack/

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.1116750.1

This is an updated version of a document previously published on Gates Open Research. 10.21955/gatesopenres.1114884.1

cabi.org | theodi.org | gatesfoundation.org

CABI Data Sharing Toolkit

open data institute

BILL & MELINDA GATES foundation