Data Sharing Toolkit

CABI

Module 6

Sharing agricultural data: managing risk to minimise harmful impacts

Guide

Sharing agricultural data: managing risk to minimise harmful impacts

About this guide

Concerns relating to data security and privacy, along with a mistrust of data or others' use of data are some of the biggest constraints to delivering data that is findable, accessible, interoperable and reusable (FAIR) within investments. Understanding risks and being able to evaluate both the real and perceived impact of data sharing is critical to overcoming these constraints by increasing confidence while minimising harmful impacts.

This guide will **help program officers** and **grantees** wanting to share data as part of an investment. Specifically it will help them to:

- Recognise where data is suitable for sharing as openly and widely as possible
- Comply with data protection and data rights legislation
- Minimise risk by identifying relevant mitigating actions open to them.

Please note this document is **not legal advice** *and if you are uncertain you should seek guidance from a legal professional.*

When to use this guide

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

Quick links

Guide: Steps to identifying and managing risk

This section takes you through three steps to identify risks in sharing data in agricultural development projects. These risk areas are generic, applicable to all data, regardless of domain, sector or geography, so these steps could be used to consider risks related to data in other domains. Within each section we suggest options to minimise the harm that could result from sharing the data.

Step 1 – Protect people

1.1 Does the data contain any personal data?

What is personal data?

Personal data is defined by the United Nations as 'information relating to an identified or identifiable natural person"'.[1]

Some types of personal data are more sensitive than others. Best practice data protection legislation (such as GDPR) defines sensitive personal information as 'special category' data and includes attributes such as race, ethnic origin, religious or philosophical beliefs, biometric data (where this is used for identification purposes) and health data.[2]

Figure 1 provides some examples of personal data and other data about people.

It is important to think about the different types of data about people and take into consideration both the opportunities, risks and the potential or perceived impact, prior to taking action. To build trust grantees and organisations using personal data should also be open with people about how they use and share that data.[3]

Figure 1. Types of data about people

[1] United Nations (2018), 'Principles on personal data protection and privacy', https://www.unsystem.org/privacy-principles. Accessed June 2020.

[2] Information Commissioner's Office UK, 'What is personal data', https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/. Accessed June 2020.

[3] Open Data Institute (2016), 'Openness principles for organisations handling personal data', https://theodi.org/article/openness-principles-for-organisations-handling-personal-data/

Countries will likely have their own definitions and categories but generally speaking any data or information directly relating to an identifiable individual is personal, including pictures of a person, or group of people. The 'agriculture data country profile' template guide in the Data Sharing Toolkit includes guidance on how to find out about local data policy.

Data protection regulations across the world are designed to minimise the risk of harmful impacts while enabling personal data to be processed, that is to be collected, accessed, used and shared.

These regulations typically outline three key things:

- The lawful basis for using and sharing personal data
- The rights of the data subject (the person the data is about)
- Liabilities and penalties for breaching the regulations.

What does personal data look like in agricultural development projects?

- Information about farm employees such as name, address or bank details.
- Mobile phone numbers – often these will have associated geo-location that could identify an individual.
- Socio-economic data such as age and education levels of each household member. This is often collected as part of project monitoring and evaluation.

- User profiles in applications that provide tailored advice to farmers may contain both personal and sensitive information.
- Free text or comments fields within the dataset – by definition these fields are not restricted in value so could easily contain personal information that might otherwise be missed. Within agricultural development projects, free text fields could be used to make notes about a relevant conversation, notes on land ownership, or contact details for a farmer, contractor or other related individuals.

If a project is accessing, using or sharing data about people, here are some ways you can minimise the risk:

Minimising harm when accessing, using or sharing data about people

- **Data minimisation.** Under many data protection regulations it is illegal to process (collect, access, use, store) personal data without a valid lawful basis. Reducing the amount of data being collected is one way of reducing risk. Ask yourself, do you really need to collect personal data? If personal details aren't vital to your research don't collect them.
- **Educate data inputters.** Educating those collecting or curating data to consider what they include in any notes within or related to the data (for example, who they spoke to) can help to avoid later problems when you want to share data. You could also design your data in a way that has no need for free text fields. Using restrictive values (e.g. date only, or pick lists) is one way to do this.

Anonymisation.[4] A technique you can apply such that data can be processed (collect, access, use, store) without needing a lawful basis as it is no longer deemed 'personal data'. Anonymisation processes data into a modified form that can be shared or made open while significantly reducing the possibility of re-identifying individuals. The anonymisation guide in the data sharing toolkit looks at how techniques can be applied to reduce the risks of re-identification and possible harm resulting from sharing data about people.

Use synthetic data. An automated process to make up (synthesise) data that contains many of the statistical patterns of an original dataset.[5] This should enable the same conclusions to be drawn from the data, but eliminate identifying personal information. This tutorial shows how to create a synthetic dataset.[6]

[4] UK Anonymisation Network, 'Definitions'. Accessed November 2020. https://msrbcel.wordpress.com/annoymisation/

[5] (2019) Open Data Institute, 'Anonymisation and synthetic data: towards trustworthy data'. Accessed June 2020. https://theodi.org/article/anonymisation-and-synthetic-data-towards-trustworthy-data/

[6] (2019), Open Data Institute, 'Anonymisation with Synthetic Data Tutorial', Accessed June 2020. https://github.com/theodi/synthetic-data-tutorial/

1.2 Does the data contain information that could adversely impact people or communities?

It is important to consider harmful broader impacts as well as legal risks – for example on individuals or communities. The increased use of data in recent decades prompts questions around issues of fairness, responsibility and accountability in relation to the use of data. It also triggers debate around whether existing legislation is fit to safeguard against harm to an individual's or group's privacy, welfare or safety.

Increasingly, those collecting, sharing and working with data are exploring the ethical implications of their practices and, in some cases, being forced to confront those implications in the face of public criticism. Thinking about the ethical use of data is particularly relevant when insights drawn or decisions informed by data have the potential to directly or indirectly impact people and communities.

When considering broader harmful impacts, think about the people the data is about, people impacted by its use and organisations using the data. For example could use of this data result in decisions that discriminate against any groups or individuals? Bias can be conscious or unconscious and can result in under-representation of specific communities, which could impact them by giving an unfair advantage to others, or unfairly restricting access (e.g exclusive arrangements), therefore it is important to consider how data collection or use might be coloured by social or personal influences.

Examples of types of information could adversely impact people or communities in agricultural projects

- An automated data model might make decisions about whether smallholder farmers are eligible for a subsidy, a financial incentive or mortgage. The decisions the model makes will be based on the data collected – and excluded (knowingly or unknowingly) – which could adversely affect groups in a society or be biased towards other groups.
- Big data is used in new technologies for 'prescriptive planting' which can automatically plant fields. Reports of mistrust relating to connections between prescriptive-planting firms collecting data and larger companies that could use the data to buy farms known to be underperforming, directly and unfairly competing with smallholder farmers.[7]
- Data published from different sources about, for example, field size and soil quality, when used together with satellite imagery, could enable competitors or financiers to obtain information they would not have historically been able to see. This could influence lease pricing or affect the ability to get loans, for example.

Consider likelihood and extent of any impact

It is important to remember that while the impact resulting from a risk could be the same for different scenarios, e.g. 'causing damage to the organisation's reputation', the likelihood and severity in each scenario could be significantly different, depending on the level of risk and the wider context.

Figure 2: Risk matrix

[7] Irish Times (2014), 'Farmers up in arms over potential misuse of data', Accessed July 2020. https://www.irishtimes.com/business/farmers-up-in-arms-over-potential-misuse-of-data-1.1863181

To assess the level of real risk of harm, it is important to evaluate the likelihood and the severity of the impact. A common method of doing this is to complete a risk matrix, like in Figure 2, and assign a risk score.

Minimising harmful impacts on people or communities

- **Consequence scanning** helps consider the intended and unintended consequences of data collection or related technologies. You can only assign risks to something if you've thought about it. Consequence scanning is part of responsible product development and can help provide structure to that thought process.
- **Consider and engage** using the Data Ethics Canvas to help you to identify and manage ethical considerations in projects involving data. It asks the user to consider 15 areas around data ethics – from bias in data sources to mitigating negative effects on people – to prompt critical thinking around how to use data ethically. It is helpful to work through the canvas as a project team to promote understanding and debate around the foundation, intention and potential impact of any piece of work, as well as help identify the steps to ensure data is handled fairly.[8]
- **Consider mitigations** for the risks, consequences or impacts that have been identified. This might include any of the steps for mitigating harm that are summarised in the previous section, such as limiting access to data or releasing only synthetic or sample data. This tutorial shows how to create a synthetic dataset.

[8] Open Data Institute (2019), 'The Data Ethics Canvas'. Accessed June 2020. https://theodi.org/article/data-ethics-canvas/

Worked example: personal data

A project is collating soil health information. The project's goals include supporting small farms with tailored advice on improving yields, and providing aggregated data to national or international monitoring programs.

Risk: Individual farms and farmers could be identified from the data.

Perceived impact: Farmers targeted or discriminated against in some way.

Minimising the risk: The project could apply anonymisation techniques to remove farmer names, use age ranges instead of date of birth and show farm location by region rather than geographical coordinates.

Result: No likely impact – anonymising the data in this way is unlikely to enable re-identification of a farmer and retains the utility of the data.

Step 2 – Protect society and the economy

2.1 Does the data contain sensitive information?

Sensitive information is any information that requires more careful handling to reduce harmful impacts. This could include considerations around privacy, politics or commercial confidentiality. Consider not just whether the data itself is sensitive but whether the data may become sensitive due to the context in which it is used.

Sensitive data can include personal and non-personal data;

- **Sensitive personal data:** Personal data that affords extra protection due to its nature, e.g. genetic data, health records or data that can potentially be used to discriminate. Figure 1 includes examples of sensitive personal data.
- **Sensitive commercial data:** Non-personal data, such as the income of a business or locations of properties, that might lead to competitors or even partners gaining a commercial advantage. It is often deemed sensitive as the use and sharing of data may still result in harmful impacts. In the context of smaller businesses, such as smallholder farmers, this type of data is possibly even more sensitive because it is likely to be largely their finance information.

What could sensitive data look like in agricultural development projects?

- Land ownership – land is a finite resource and there is competition to control and exploit it. Effective access to land data for one user may lead to significant first-mover's advantage and therefore preclude other users from taking action relating to a parcel of land, even if they eventually have access to the same data.[9] Equally important to consider are the risks of not sharing data. For example not sharing land ownership data may cause more discrimination and bias towards some actors than sharing it. There needs to be a balance here and this is where the risk matrix and ethics canvas can help.
- Plant, pest and disease data can be sensitive when considered in the light of crop exports and their trade implications. .
- Legal enforcement such as data detailing the compliance of organisations with particular regulations and laws, or details of legal action pending or taken where it might influence business or financial decisions.
- Agri-supply data related to use and composition of pesticides, fertilizers, feed. Some pesticides, fertilizers and feed might be classified as hazardous materials that could damage the environment, for example by leaching into the water supply.

[9] Open Data for Development (2019), 'State of Open Data', Accessed June 2020. https://stateofopendata.od4d.net/chapters/sectors/land-ownership.html

2.2 Does the data contain anything confidential?

Confidential information includes anything that requires restricted access. Usually this will be decided by the organisation who created or stewards that information. It could relate to commercial activities or it could also be personal or sensitive information.

In agricultural development projects confidential data could include financial figures such as subsidies, crop prices, farm income or pesticide formulas that a farmer or organisation does not wish anyone else to be able to access.

2.3 Does the data contain any information that could impact national security?

National security, is broadly defined as the safety of a nation against threats such as terrorism, war, natural disaster, and could be put at risk through the release of data. This includes any data that could be used to cause actual harm, deprivation or fear of the same.

For example, releasing information that enables the identification of key drinking water and agricultural irrigation sources , or of key food production and storage areas for internal use and export could be a target for terrorist activities or in times of conflict.

Minimising harm when collecting, accessing, using or sharing sensitive or confidential data, or data that could impact national security

- **Educate data inputters** by making those collecting or curating data aware of the types of data that might need particular attention and to consider what they include in any notes within or related to the data (e.g. informal notes on the location of pesticides on site). It can also help to design your data in a way that has no need for free text fields and instead use restrictive values (e.g. date only, or pick lists) to prevent sensitive data being included inadvertently.
- **Suppression.** Processing data into a modified form that conceals certain elements of the data set, so that it can be shared or made open while significantly reducing the possibility of anyone recovering sensitive information from it.
- **Use synthetic data.** Using an automated process to make up (synthesise) data that contains many of the statistical patterns of an original dataset.[10] This should enable the same conclusions to be drawn from the data, but eliminate any information that is sensitive or confidential in nature, or that could pose a security risk. This tutorial shows how to create a synthetic dataset.[11]
- **Share the data under contract.** Data sharing agreements can be useful when organisations of any kind are sharing data that is commercially confidential or includes information of a sensitive nature. A contract with detailed, binding rules ensures all parties are clear on their obligations. The guide on designing data sharing agreements includes a checklist that can help you create an effective contract.[12]

[10] (2019) Open Data Institute, 'Anonymisation and synthetic data: towards trustworthy data'. Accessed June 2020. https://theodi.org/article/anonymisation-and-synthetic-data-towards-trustworthy-data/

[11] (2019), Open Data Institute, 'Anonymisation with Synthetic Data Tutorial', Accessed June 2020. https://github.com/theodi/synthetic-data-tutorial/

[12] Deborah Yates, Tim Beale, Stewart Marshall, Martin Parr (2018), 'Designing data sharing agreements: a checklist', Accessed July 2020. https://gatesopenresearch.org/documents/2-44

Worked example: national security

A transport authority wants to release the live locations of trains informing public travel and commercial goods transport. The dataset includes freight trains carrying nuclear waste.

Risk: Terrorists could use the data to target trains.

Perceived impact: Damage to trains and railway infrastructure, possible harm to passengers and staff.

Minimising harm: Using a combination of suppression and synthetic data to remove the names and actual identifiers of freight trains and creating variable identifiers in their place will reduce the likelihood of individual freight trains being tracked and targeted.

Result: Nothing yet for freight trains but people did use the data to follow “Flying Scotsman” (steam engine) and some trespassed on the track as a result causing huge delays. Special services hauled by “Flying Scotsman” are now given synthetic and variable identifiers to mask them.

2.4 Does the data impact any third party rights or contain data created by someone else?

When an individual or as an organisation puts intellectual effort into creating something, such as taking a photograph or collecting data, the law grants you specific rights of ownership over that work.

Countries will likely have their own laws and definitions but generally speaking, by default, the data creator holds exclusive rights to use the data, so that 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 to ensure the permissions that are acquired facilitate the onward licensing and use of data.

Grantees might be making use of a range of data sources, and not have permission to use or share all the data available or used within an investment. In particular the resources might:

- Be completely licensed from someone else
- Include an extract of content or data licensed from someone else
- Be derived from the content or data licensed from someone else.

What does third party data look like in agricultural development projects?

The Data Sharing Toolkit guide 'considering data rights and permissions in investments' provides more detail on this topic, however here are a few examples of common types of data that has been sourced from third parties for use within investments:

- Earth observation data – Data from satellites is often used for large-scale harvesting of environmental information. For example, NASA's earthview, google maps or the European Space Agency maps.
- Climatic or weather data – Data from organisations that monitor, record or predict the weather. For example, monthly temperatures, day and night time, and monthly rainfall. Often it is possible to access weather forecasts online, but we do not necessarily have the right to use the data.[13]
- Government data – Data about a country, its regions, infrastructure and populations as collected by national and local governments. For example, a population census, land boundaries, political or administrative boundaries, regulatory compliance data, settlements and road networks.
- Research data – Data collected as part of academic, commercial or private research. For example, samples such as soil PH, air particle concentration, or water salinity.

[13] Open Knowledge Foundation, Global Open Data Index – weather forecast, Accessed July 2020. https://index.okfn.org/dataset/weather/

Minimising harm when sharing or using data that has been collected or created by a third party

- Discuss with **third party stewards.** Ensure you have permission to use data by convening a conversation with the third party data steward. This can help to overcome any possible Intellectual Property (IP) issues and establish that onward use and sharing won't cause any harm. If the data is not under an open licence and the steward places some restrictions on how the data can be used or shared, it may be possible to agree to share the data under a contract or licence that complies with the relevant permissions.
- **Cover all bases.** In some cases, the reuse permissions related to a dataset might be unclear and a rights holder not listed. Investigate as far as possible until you are sure all your options have been exhausted, including any fair use (or fair dealing) exceptions or other local laws that might cover the use of the data. If this is not the case, then a risk-based decision could be made over the collection, use and sharing of the data, with the understanding that the rights holder could emerge and it may be necessary to negotiate continued use or potentially face a legal challenge.
- **Consult the Data Sharing Toolkit** which contains further relevant guides that will be useful on this topic, including considering data rights and permissions in investments and How to choose an open data licence.

Step 3 – Encourage best practice

3.1 Is the data properly described and documented to protect from misuse?

One of the core constraints to delivering FAIR and safeguarded data within investments is a mistrust in others' use of data. Concerns around misuse of data can include:

- Drawing 'incorrect' conclusions that might be attributed to the data publisher
- Exposing issues with the data publisher's operations due to limitations in the data
- Harming the business activities of the publisher, for example by allowing someone to replicate a service or product

Concerns that others would misuse data often stem from the assumption that others do not understand enough about the data. This can often lead to data not being shared. However, it is also important to ask how you battle misrepresentation or misuse? Any organisation could publish an analysis that you know to be false, regardless of whether the data is available or not, however people cannot refute the claims of others unless the data is available for them to analyse themselves. Sharing data as openly as possible can in fact fight misuse and misrepresentation rather than lead to more of it!

Minimising harms and risk through good data documentation

Describing and documenting data is best practice and will ensure reusers understand important context, e.g. how it was collected, for what purpose and known limitations with the data. Data documentation can help users to understand whether they might be able to use it and also help to manage concerns around mis-use of data. This guide on describing and documenting data well should help you to do this.

Seeking further help

Identifying risk and minimising harmful impacts from sharing data can sometimes require a specialist team including:

- Domain or Policy specialists to inform what is required to be shared, how the data is to be used as well as the validity of data exchanged.
- Data and information specialists that understand the technical aspects of the data to be shared and how the data may be integrated with other data sources that could raise other legal issues not inherent within the data alone.
- Intellectual property rights (IPR) experts
- Legal support from Bill & Melinda Gates Foundation lawyers, and in some instances external legal guidance.

You may need to consult colleagues, partners and specialists when considering risks and how to minimise harm.

In all cases, this document is **not legal advice** and if you are uncertain you should seek guidance from a legal professional.

Appendix: Checklist to identify and manage risk when sharing data

This checklist summarises key areas of risk to consider when sharing data. The questions aim to help minimise the harmful impacts of data being collected, accessed, used and shared within an investment, while ensuring it is shared as widely as possible.

Using the checklist

1. Answer all questions in steps 1 and 2
    a. If you answer ‘no’ to any of the questions it is unlikely the data contains anything that could harm people. The data can be considered low risk. Sharing it as widely as possible is unlikely to result in harm to individuals or society. Move to step 3.
    b. If you answer ‘yes’ or ‘unsure’, then move to the relevant sections in this guide, or supporting eLearning in the FAIR data toolbox, to consider options to minimise the risk of harm. Move to step 3.

2. Answer all question in step 3
    a. If you answer ‘yes’ the risk of harm is likely to be low however we still encourage best practice in data curation and documentation. Move to the relevant section in this guide to find out more.
    b. If you answer ‘no’ or ‘unsure’ then move to the relevant section in this guide to to consider options to minimise the risk of harm.

If you are still unsure or you have identified risks of harm in need of mitigation, **speak to a legal specialist.**

Step	Question	Yes	No	Unsure
Step 1				
Protect people	1.1 Does the data contain any personal data?			
	1.2 Does the data contain anything that could adversely impact people?			
Step 2				
Protect society and the economy	2.1 Does the data contain anything sensitive?			
	2.2 Does the data contain anything confidential?			
	2.3 Does the data contain anything that could impact national security?			
	2.4 Does the data contain any third party rights?			
Step 3				
Encourage good practice	3.1 Is the data properly described and documented to protect from misuse?			

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.1116754.1

cabi.org | theodi.org | gatesfoundation.org

CABI Data Sharing Toolkit

open data institute

BILL & MELINDA GATES foundation